Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Ciência da Computação

Geoprocessamento e mineração de dados na identificação de riscos e ganhos no faturamento e no suporte a decisão de expansão

Carlos Eduardo Machado Pires

Dissertação apresentada como requisito parcial para conclusão do
Mestrado Profissional em Computação Aplicada

Orientador
Prof. Dr. João Mello da Silva

Coorientador
Prof. Dr. Marcelo Ladeira

Brasília
2019

Ficha catalográfica elaborada automaticamente,
com os dados fornecidos pelo(a) autor(a)

Mg

Machado Pires, Carlos Eduardo
Geoprocessamento e mineração de dados na identificação de riscos e ganhos no faturamento e no suporte a decisão de expansão / Carlos Eduardo Machado Pires; orientador João Mello da Silva; co-orientador Marcelo Ladeira. -- Brasília, 2019.
92 p.

Dissertação (Mestrado - Mestrado Profissional em Computação Aplicada) -- Universidade de Brasília, 2019.

1. Mineração de Dados. 2. Riscos ao faturamento. 3. Geoprocessamento. I. da Silva, João Mello, orient. II. Ladeira, Marcelo, co-orient. III. Título.

Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Ciência da Computação

Geoprocessamento e mineração de dados na identificação de riscos e ganhos no faturamento e no suporte a decisão de expansão

Carlos Eduardo Machado Pires

Dissertação apresentada como requisito parcial para conclusão do
Mestrado Profissional em Computação Aplicada

Prof. Dr. João Mello da Silva (Orientador)
PPCA/UnB

Prof. Dr. Edgard Costa Oliveira
PPCA/UnB

Prof. Dr. Sérgio da Costa Cortês
IESB

Prof.a Dr.a Aletéia Patrícia Favacho de Araújo
Coordenadora do Programa de Pós-graduação em Computação Aplicada

Brasília, 31 de julho de 2019

Dedicatória

Dedico este trabalho ao ser mais especial e importante da minha vida.

Yuri, meu filho, que o empenho e dedicação de desprendi neste projeto de vida, lhe seja fonte de inspiração para acreditar em sí mesmo, nos seus sonhos e catalizar sua força de vontade e dedicação para conquistar tudo o que almejar.

Agradecimentos

Chegar até aqui foi uma longa caminhada. Uma trilha que não se pecorre sozinho... Diversas pessoas participaram, direta ou indiretamente desta jornada, compartilhando das angustias, cedendo paciência e compreensão, apoio e orientação. Pontanto, neste momento de colheita, não tem como não se sentir grato e lembrar, dedicar os resultados a cada uma delas.

Agradeço a DEUS, que abriu portas, me guiou e colocou cada uma dessas pessoas na minha vida.

Agradeço ao amigo/irmão Eduardo Sousa que me convidou, incentivou e desafiou a cursar este mestrado, me tirando da zona de conforto e instigando a continuar na necessária e engrandecedora busca pelo conhecimento.

Agradeço aos ilustres professores do PPCA, em especial aos meus orientadores Prof. Dr. João Mello e Marcelo Ladeira, por compartilhar comigo uma pequena parte de sua enorme gama de conhecimento. Mestre é aquele que acolhe nossas limitações com a ternura das palavras e aponta o caminho com a firmeza das atitudes. Obrigado pelos ensinamentos, orientação e condução deste projeto.

Registro, ainda, o apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior do Brasil (CAPES), por possibilitar acesso ao portal de periódicos.

Agradeço enormemente aos meus colegas de trabalho, em especial à excepcional equipe da Gerência de Geoprocessamento que muito contribuíram na parte prática desta pesquisa e aos membros do Grupo de Trabalho da DT 396/2018, que acreditaram neste projeto, dispuseram de seu tempo e conhecimento para fornecer informações fundamentais para condução da pesquisa e a equipe técnica do grupo de trabalho que se dedicou com comprometimento ímpar na etapa de análise dos dados.

Por fim, meu especial agradecimento à Márcia Sabino Duarte, amiga de longa data e agora, como um presente divino, minha companheira e meu porto seguro. Obrigado pelas palavras e atitudes motivadoras, pelo apoio nos momentos difíceis e por, muitas vezes, acreditar em mim mais do que eu mesmo...

Resumo

Segundo o Sistema Nacional de Informação sobre Saneamento - SNIS, em 2016 a perda de faturamento da Companhia de Saneamento Ambiental do Distrito Federal, Caesb, foi de 24,71% decorrentes das perdas de água.

As perdas de água são, majoritariamente, combatidas por ações de engenharia para detecção e correção de vazamentos nas redes, as chamadas perdas reais. No entanto, as perdas denominadas aparentes (aquelas em que a água produzida é consumida mas não faturada) são de difícil detecção, de modo que seu combate é realizado por investigações em campo após denúncias de fraudes ou seleção aleatória ou pouco sistematizada dos locais a serem investigados.

Neste contexto, a mineração de dados se mostra ferramenta fundamental na identificação de riscos ao faturamento possibilitando otimizar as investigações *in loco* e descoberta de possibilidades de ampliação de receita.

Esta pesquisa apresenta o estudo de caso da Caesb, onde duas abordagens foram adotadas visando identificar, por meio da mineração de dados, os potenciais riscos ao faturamento da companhia.

A primeira abordagem focou na criação de um modelo preditivo para identificação de potenciais fraudes no consumo de água, onde duas hipóteses foram elaboradas, testadas e refutadas por problemas de consistência nos dados.

A segunda abordagem pautou-se no conceito de par perfeito. Este conceito parte da premissa que o cliente deve gerar receita para Caesb pelos serviços de abastecimento de água e coleta de esgoto.

Com este pressuposto, foi elaborada, testada e confirmada a hipótese de que clientes localizados em regiões em que existem redes de abastecimento de água e/ou esgotamento sanitário mas que não pagam por quaisquer destes serviços resultam em potencial risco ao faturamento da companhia.

Neste sentido, o modelo de mineração de dados criado combina dados comerciais e geoespaciais para descoberta dos clientes que violam o conceito de par perfeito e, por meio de análise geoespacial, realiza a clusterização destes clientes em quatro grupos dis-

tintos: inconsistência cadastral, problema operacional, problema de expansão e problema de extensão.

O modelo de mineração de dados é executado mensalmente, de forma automática, a cada fechamento comercial e o resultado da mineração apresentado em um *dashboard* web que possibilita identifica a quantidade, localidade e categoria dos imóveis que violam o par perfeito, bem como estimar o impacto financeiro causado.

Em maio de 2019, o modelo detectou 119.887 potenciais situações que geram impacto financeiro, entre perda de faturamento e oportunidade de aumento de receita, na ordem de R$ 120 Milhões/ano, ou seja, 7,5% do faturamento anual da empresa.

Palavras-chave: Perdas Financeiras, Consumo, Água, Mineração de Dados, Riscos, Geoprocessamento

Abstract

According to the National Sanitation Information System - SNIS, in 2016 the loss of revenues of the Environmental Sanitation Company of the Federal District, Caesb, was 24.71% due to water losses.

Water losses are mostly countered by engineering actions to detect and correct leaks in the networks, the so-called real losses.

However, the so-called apparent losses (those in which the produced water is consumed but not billed) are difficult to detect, so that their combat is carried out by field investigations after allegations of fraud or random or poorly selected sites to be investigated.

In this context, data mining proves to be a fundamental tool in identifying billing risks, enabling the optimization of on-site investigations and the discovery of revenue expansion possibilities.

This research presents Caesb's case study, where two approaches were adopted to identify, through data mining, the potential risks to the company's revenue.

The first approach focused on the creation of a predictive model for identifying potential water consumption frauds, where two hypotheses were elaborated, tested and refuted by data consistency problems.

The second approach was based on the concept of perfect match. This concept assumes that the client must generate revenue for Caesb for water supply and sewage collection services.

Based on this assumption, it was developed, tested and confirmed that customers located in regions where there are water supply and/or sewage systems but who do not pay for any of these services result in a potential risk to the company's revenues.

In this sense, the data mining model created combines business and geospatial data to discover clients that violate the perfect pair concept and, through geospatial analysis, cluster these clients into four distinct groups (cadastral inconsistency, operational problem, expansion problem and extension problem).

The data mining model is automatically executed monthly at each trade close and the mining result presented in a web dashboard that identifies the number, location and

category of properties that violate the perfect match, as well as estimating the impact caused.

In May 2019, the model detected, 119,887 potential situations that generate financial impact were detected, such as loss of revenue and opportunity to increase revenue, in the order of R$ 120 million / year, or 7.5% of the annual revenue of the Company.

Keywords: *Financial loss, Water Consumption, Data Mining, Risk, Geoprocessing*

Sumário

Lista de Figuras

Lista de Tabelas

Capítulo 1

Definição do Problema

A Companhia de Saneamento Ambiental do Distrito Federal - Caesb é responsável pela captação, tratamento e distribuição de água tratada para toda a população do Distrito Federal, que nos últimos 26 anos cresceu 87,5%, saindo de uma população de 1 milhão e 600 mil pessoas em 1991 para pouco mais de 3 milhões em 2017 [2] [3].

Este acentuado crescimento populacional, além das ocupações irregulares que provocam desmatamento predatório, impermeabilização do solo, assoreamento de nascentes e mananciais e aumento da captação de água para produção agrícola reduziram a vazão média de afluente, ou seja, a capacidade de recomposição dos reservatórios, que somados com a baixa história de precipitações, contribuíram para uma diminuição nos níveis dos principais reservatórios do Distrito Federal como nunca antes observado [1].

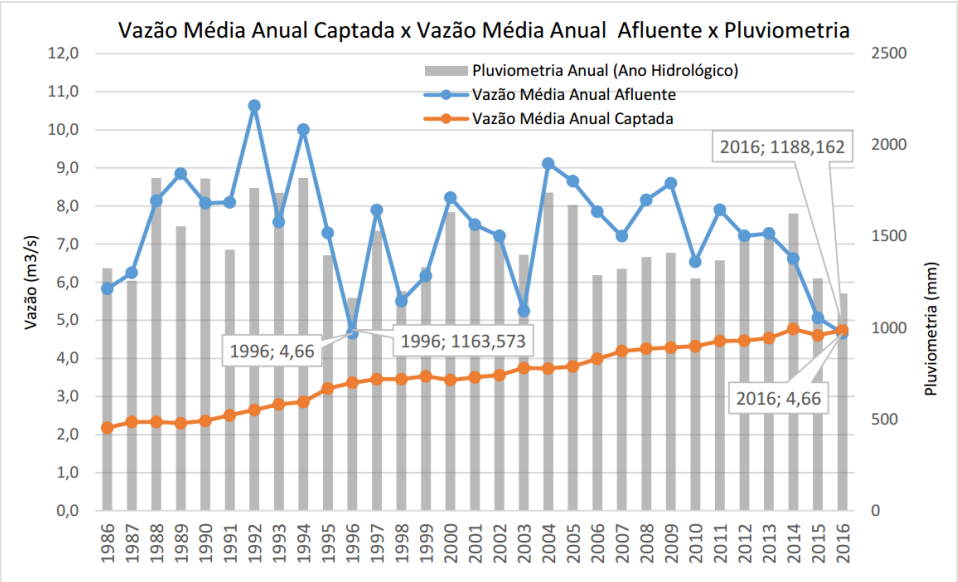


Figura 1.1: Histórico de recomposição e captação na Barragem Descoberto [1]

A Figura 1.1 apresenta uma comparação histórica entre a capacidade de recomposição do reservatório do Descoberto, responsável pelo fornecimento de água para 61,52% da população do Distrito Federal [1]. Neste gráfico fica evidente que o crescimento populacional resultou em um expressivo aumento da vazão de captação, enquanto a capacidade de recomposição do reservatório, representada pela precipitação e vazão de afluente, tiveram uma sensível redução. Isso resultou na grave crise hídrica de 2017, onde foi necessário implantar um plano emergencial de racionamento, deixando a população do DF sem abastecimento de água por pelo menos um dia por semana [4].

A escassez hídrica não é um problema isolado no Distrito Federal. A comissão mundial para desenvolvimento e meio ambiente afirma que 80 países ao redor do mundo, com cerca de 40% da população mundial, sofrem com sérios problemas de escassez [5] e, segundo a ONU, diversos países têm cada vez mais se preocupado e procurado formas de fazer exploração sustentável dos recursos hídricos [6].

Neste sentido, além de ações emergenciais e construções de novas fontes de captação, o Governo do Distrito Federal tem realizado intensas campanhas de conscientização, incentivando a população a um uso racional da água.

Não obstante, ações estratégicas e operacionais também devem ser adotadas para minimizar os impactos da crise hídrica, sendo o combate as perdas de água a mais relevante destas ações, uma vez que as perdas que ocorrem no sistema de abastecimento de água é o mais grave problema no setor de saneamento público pois afetam a sustentabilidade do abastecimento e impactam negativamente o meio ambiente, em especial os escassos recursos hídricos, já que é necessário aumentar o volume de captação dos mananciais para atendimento as demandas de consumo [7][8][9].

O Sistema Nacional de Informações sobre Saneamento - SNIS, apresenta anualmente no diagnóstico de serviços de água e esgoto no Brasil. Dentre as informações publicadas neste diagnóstico, destaca-se o índice de perda de água que é calculado considerando a seguinte fórmula:

$$\frac{(VolumeProduzido - VolumeConsumido)}{VolumeProduzido} - VolumeOperacional \tag{1.1}$$

O diagnóstico de serviços de 2016, publicado em fevereiro de 2018, revela que em média 37% de toda água coletada, tratada e distribuída no Brasil são perdidas seja com vazamentos, ligações clandestinas, fraudes ou problemas na medição do consumo [1]. Esta perda, além do impacto ambiental, resulta em prejuízos financeiros as concessionarias de serviços de abastecimento público em razão do volume tratado, distribuído e não faturado [8][7].

[1]Fonte: http://www.snis.gov.br/diagnostico-agua-e-esgotos/diagnostico-ae-2016

Os indicadores operacionais da Caesb, publicadas neste mesmo diagnóstico de serviços do SNIS, demonstra que o índice de consumo de água no Distrito Federal é de 64,79%, ou seja, 35,21% da água coletada e tratada é perdida no sistema de distribuição. Este índice representa uma perda de faturamento de 24,71%, ainda segundo o SNIS.

As perdas de água no sistema de abastecimento se dividem em duas categorias [8][9][10]:

1. **Perdas Reais**, também conhecidas como perdas físicas, referem-se às perdas de água motivadas por vazamentos nas tubulações de adução ou distribuição da água coletada e tratada.

2. **Perdas Aparentes**, ou perdas comerciais, referem-se ao volume de água que foi efetivamente consumido pelo usuário mas não foi medido ou contabilizado, gerando perda de faturamento. Estas perdas são motivadas por falhas na medição (problemas nos dados ou hidrômetros ineficientes) ou consumo não autorizado (fraudes ou ligações clandestinas).

As melhores práticas no gerenciamento de perdas de água consiste no contínuo monitoramento do balanço hídrico que quantifica o volume total de água produzida para abastecimento, o consumo autorizado e a perda de água (Figura 1.2)[9]

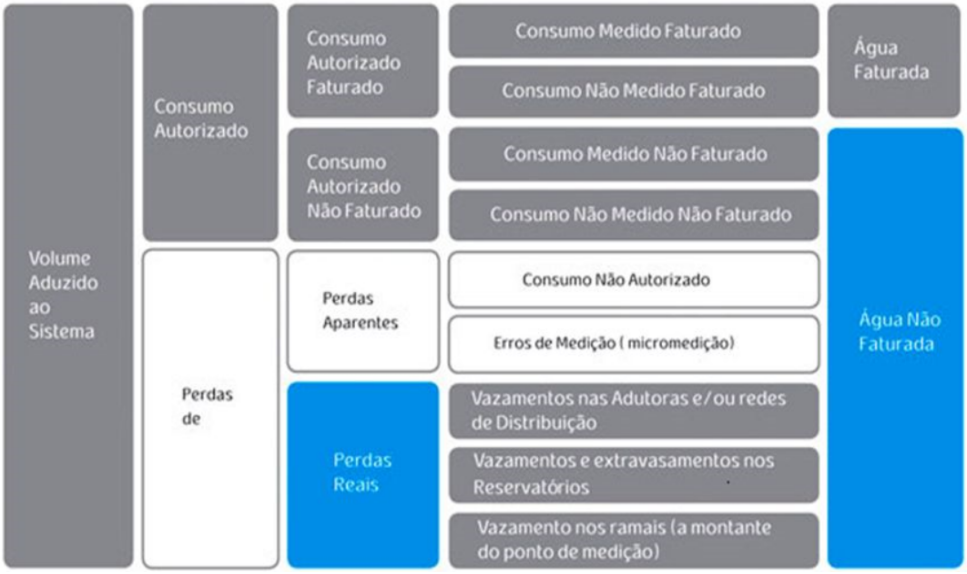


Figura 1.2: Componentes do Balanço Hídrico
Adaptado do IWA

A Caesb aplica as orientações da IWA para realizar o monitoramento periódico da produção e consumo de água tratada no Distrito Federal, visando identificar e quantificar as perdas reais e aparentes de modo a possibilitar o planejamento de ações estratégicas e corretivas para redução destas perdas (Figura 1.3).

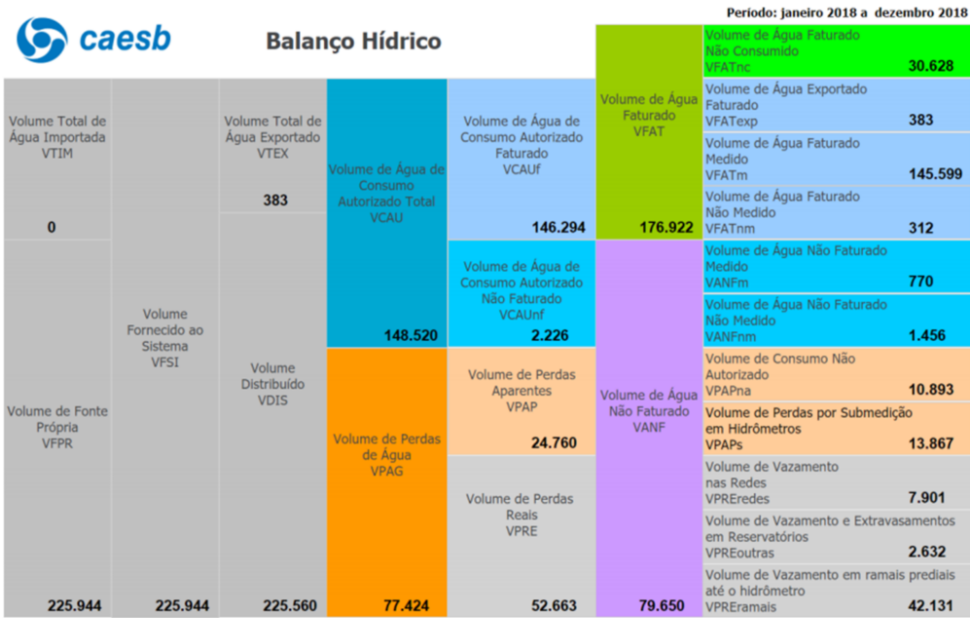


Figura 1.3: Balanço Hídrico do DF/2018

Percebe-se, pela Figura 1.3, que as perdas de água, sejam aparentes ou reais, é responsável por grande parcela do consumo de água não faturado de modo que, combatendo-se as causas da perda de água, reduz-se o prejuízo financeiro da companhia e, como consequência indireta, melhora-se o serviço prestado por alcançar maior eficiência operacional [7] o que reflete na tarifa cobrada dos usuários, uma vez que os custos das perdas de água, invariavelmente, são repassados ao consumidor [8] e minimiza os impactos ambientais, uma vez que reduz-se a captação para atendimento à demanda da população.

Enquanto o combate as perdas reais é realizado, principalmente, por ações de engenharia como obras de revitalização das tubulações, modernização do sistema de abastecimento e monitoramento de setores pra detecção de vazamentos, o combate as perdas aparentes demanda a utilização de ferramentas de tecnologia da informação para identificar, mitigar ou até mesmo eliminar fatores que impactam no faturamento [7].

1.1 Justificativa

Diferente das perdas reais, onde os vazamentos podem ser detectados com a utilização de equipamentos acústicos, as perdas aparentes são mais complexas de serem identificadas pela diversidade de variáveis na definição de um padrão de consumo, como alteração na quantidade de moradores, viagens, clima, vazamentos entre outros [7]. Obviamente que uma variação significativa no padrão de consumo de uma unidade é um indício de

fraude, demandando ações de vistorias em campo para efetivamente identificar e sanar as irregularidades.

No entanto, em geral, as investigações de fraudes são executadas de forma aleatória pelas concessionárias de serviço público de abastecimento [7]. O mesmo ocorre na Caesb, onde as vistorias para identificação de fraudes são realizadas, majoritariamente, sem fundamentação sistematizada para detecção de possíveis unidades de consumo com irregularidade.

A adoção de métodos e sistemas para identificação de fraudes e outros riscos ao faturamento elimina as dificuldades e aleatoriedade nas investigações em campo, direcionando-as para as unidades consumidoras com maior probabilidade de práticas irregulares no consumo de água.

Neste sentido, a mineração de dados tem se mostrado eficaz em diversas áreas do conhecimento e diferentes setores produtivos, como por exemplo empresas de telecomunicação, companhias financeiras e agências governamentais, para descoberta de informações relevantes que proporcionam o ganho de receitas e redução de perdas financeiras [11].

Considerando os casos de sucesso relatado acima em situação-problema similiar ao da Caesb, a presente pesquisa propõe um modelo de mineração de dados que auxilia na identificação precisa de fatores com potencial risco ao faturamento da companhia.

1.2 Objetivos

1.2.1 Objetivo Geral

Este trabalho tem por objetivo principal identificar, por meio da análise de dados comerciais da Caesb, clientes que potencialmente possam impactar negativamente no faturamento da companhia e\ou possibilidades de aumento de receita.

1.2.2 Objetivos Específicos

Para atingir o objetivo geral, a presente pesquisa tem como objetivos específicos:

1. Identificar inconsistências cadastrais que resultam em perda de faturamento;
2. Identificar clientes que violam o conceito de par perfeito e impactam no faturamento da companhia;
3. Auxiliar na tomada de decisão de ações de expansão de redes considerando potencial de retorno financeiro;

4. Implementar mapa dinâmico para destacar clientes que impactam no faturamento e oportunidades de aumento de receita.

1.3 Contribuição Esperada

Ao atingir o objetivo geral, espera-se que este trabalho contribua na minimização dos custos operacionais para distribuição de água e atendimento as ações necessárias para atingir a meta 6.4 da ODS 6 (Capítulo 3.1.2), uma vez que a identificação de riscos ao faturamento possibilitará executar ações para garantir a redução das perdas financeiras, ao mesmo tempo que se otimiza o processo de vistoria, aumentando a produtividade da equipe com a redução de visitas improdutivas, ou seja, vistorias em locais em que não existem riscos ao faturamento da companhia.

Atingindo os objetivos específicos, por sua vez, espera-se possibilitar à Caesb melhorar a qualidade do cadastro comercial no que tange à situação da ligação de água e esgoto, além de permitir o monitoramento das ações saneadoras dos clientes que violam o par perfeito e apoiar nas tomadas de decisões estratégicas para investimento em expansão das redes de abastecimento de água e esgotamento sanitário.

Capítulo 2

Metodologia de Pesquisa

Para atingir os objetivos desta pesquisa foram adotadas duas linhas de trabalho, conforme demonstrado na Figura 2.1: i) pesquisa exploratória na literatura especializada para identificar padrões, tendências e aplicações práticas relacionadas ao tema de fundo desta pesquisa para criação de um modelo de identificação de riscos ao faturamento e ii) estudo de caso, onde o modelo criado foi aplicado aos dados da Caesb.

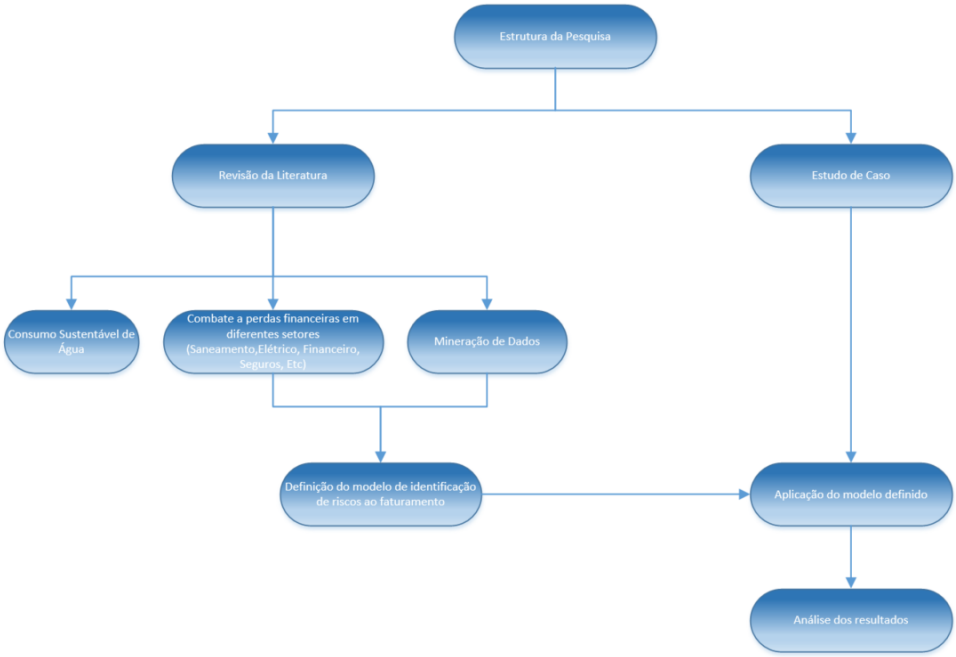
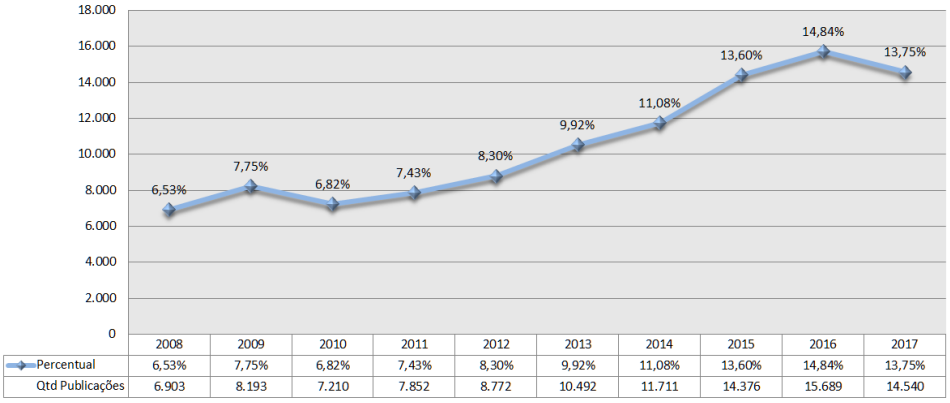


Figura 2.1: Estrutura da pesquisa

Por fim, os resultados gerados pela aplicação do modelo foram analisados para certificar se os objetivos específicos foram atingidos.

2.1 Parâmetros de pesquisa da revisão da literatura

O tema de fundo deste trabalho, que utiliza métodos e técnicas de mineração de dados (*data mining*) para identificação de riscos ao faturamento da Caesb, apesar de não ser algo novo, está em notável expansão de modo que a literatura sobre o assunto é vasta e abrangente. O portal *Web of Science*, uma das bases de dados de referência internacional em publicações científica, demonstra um crescimento de 110% nas pesquisas em *Data Mining* nos últimos 10 anos (Figura 2.2).



	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017
Percentual	6,53%	7,75%	6,82%	7,43%	8,30%	9,92%	11,08%	13,60%	14,84%	13,75%
Qtd Publicações	6.903	8.193	7.210	7.852	8.772	10.492	11.711	14.376	15.689	14.540

Figura 2.2: Publicações por Ano.

Ao longo da última década, identificou-se um total de 105.738 trabalhos científicos relativos à mineração de dados em diversas áreas de pesquisa (Figura 2.3).

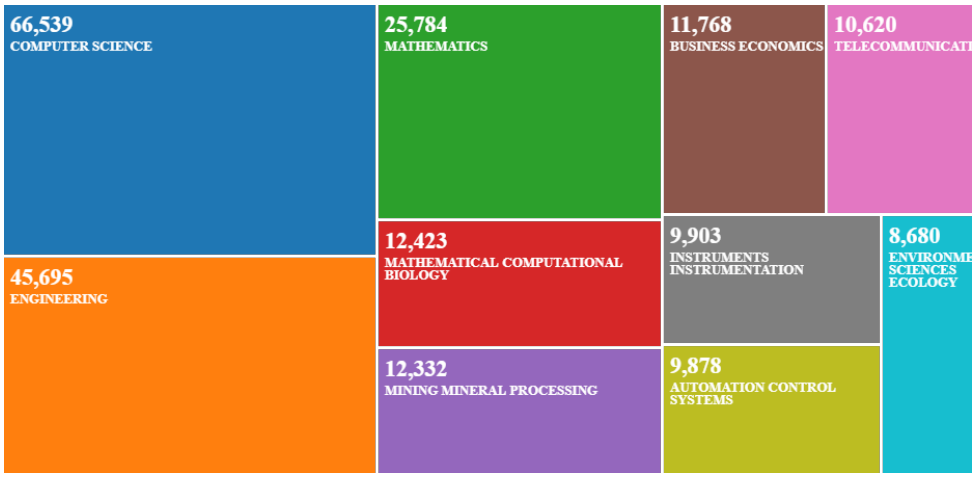


Figura 2.3: Publicações por Área de Pesquisa.

A gama de opções bibliográficas dificulta a identificação de publicações relevantes e representativas sobre a temática desta pesquisa. Para sanar esta dificuldade, existem

métodos de pesquisa acadêmica que auxiliam no mapeamento do conhecimento científico e na busca de literatura relevante para a pesquisa, dos quais se destaca a Teoria do Enfoque Meta Analítico - TEMAC [12].

Esta técnica não objetiva estabelecer limites à bibliografia que deve ser consultada, pelo contrário, demonstra publicações relevantes e de impacto que muito contribuirão ao trabalho em andamento, servindo como ponto de partida para a revisão da literatura. Portanto, outras fontes bibliográficas podem e devem ser consultadas para complementar o embasamento teórico da pesquisa.

A TEMAC sugere a execução de três etapas na busca de publicações relevantes: 1) preparação da pesquisa com a definição de parâmetros como palavras chaves, base de dados a ser consultada e período e local das publicações; 2) apresentação e inter-relação dos dados que permite identificar os autores e periódicos de maior impacto e 3) detalhamento e validação por evidências que apontam as abordagens referentes ao tema de estudo com maior relevância.

Na primeira etapa, a pesquisa exploratória foi realizada nas bases de trabalhos científicas consolidadas e internacionalmente reconhecidas como relevante e de importante contribuição para a ciência [12]: ISI Web of Science-WoS (http://www.webofknowledge.com), Scopus (http://www.scopus.com) e Google Scholar (http://scholar.google.com), onde se buscou por trabalhos publicados na área de pesquisa da computação.

As palavras chaves utilizadas considerou os seguintes escopos de pesquisa:

1. **Modelos de identificação de riscos ao faturamento em setores diversos**, as fraudes são os principais fatores de riscos ao faturamento de qualquer empreendimento, portanto, utilizou-se as palavras chaves *fraud detection* para conhecer os principais modelos de detecção de fraudes aplicados em situações reais em diversos setores da economia, como energia elétrica, consumo de água, operações financeiras, cartões de crédito e seguradoras de saúde;

2. **Métodos e Técnicas de Mineração de Dados**, ao perceber que a mineração de dados foi a abordagem mais utilizada para detecção de fraudes, realizamos pesquisas nos métodos e técnicas identificados na etapa anterior por meio das palavras chaves *datamining, data mining, fraud detecction, thecniques, methodos* para localizar publicações que se aprofundam nos conceitos, métodos e técnicas de mineração de dados aplicados à detecção de fraudes.

A fundamentação teórica obtida por meio da revisão da literatura apoiada pelo enfoque meta analítico, que por ser uma técnica objetiva que respalda a escolha da literatura [12], subsidiará na definição de um modelo para proceder com a análise de dados da Companhia de Saneamento Ambiental do Distrito Federal visando identificar fatores que impactem no

faturamento da companhia de modo a apoiar na definição de planos de ação que permita à companhia minimizar as ocorrências destes fatores.

2.2 Fontes de pesquisa relevantes

Os parâmetros de pesquisas apresentados na subseção anterior foram aplicados em cada um dos escopos de pesquisa deste trabalho onde foi possível, por meio de técnicas de bibliometria, identificar os autores e publicações de maior impacto em cada temática.

A busca por bibliografia no escopo de pesquisa de Métodos e Técnicas de Mineração de Dados, utilizando a palavra chave *data mining* por exemplo, resultou em 1.628 publicações. A análise de co-citação nestas publicações, considerando autores que foram citados em conjunto mais de 20 vezes, resultou no mapa de calor da Figura 2.4, que revela três autores (Agrawal, Fayyad e Han) cuja abordagem tem grande influência nos trabalhos atuais.

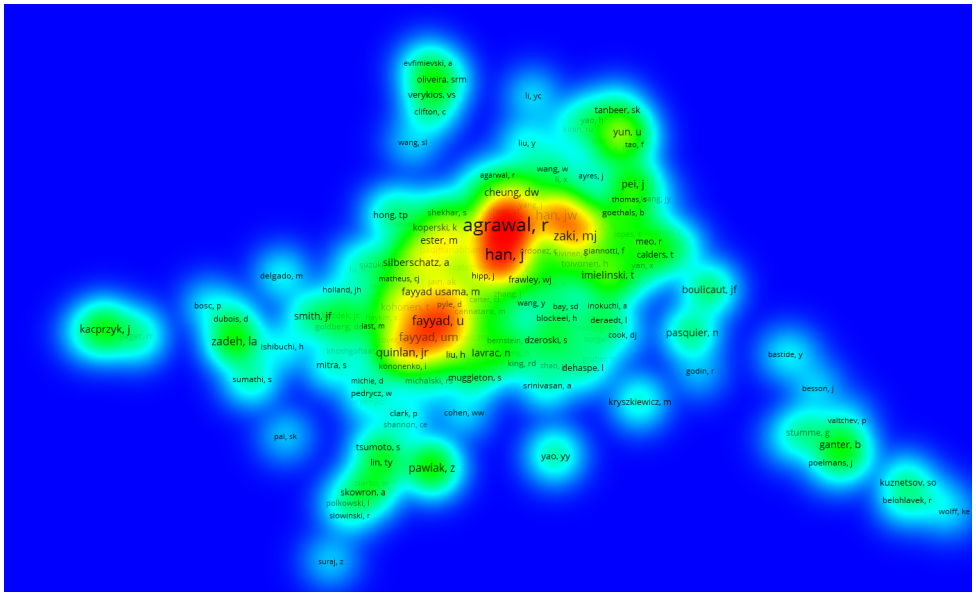


Figura 2.4: Autores influentes no tema Data Mining

Por este motivo, em que pese suas publicações não serem recentes, estes autores foram considerados como ponto de partida na revisão da literatura neste escopo de pesquisa.

Estes procedimentos foram adotados nos demais escopos de pesquisa, revelando as publicações de maior influência que foram consultados para fundamentação da base teórica da pesquisa. Os resultados foram omitidos por serem semelhantes, no que tange à apresentação, ao demonstrado acima.

Capítulo 3

Revisão da Literatura

3.1 Água Potável: Ações para consumo sustentável

Como exposto na introdução deste trabalho, o Distrito Federal, assim como diversas outras cidades ao redor do mundo, vem enfrentando uma escassez hídrica sem precedente.

Neste contexto, a presente sessão apresenta, em linhas gerais, ações adotadas por organizações com presença e influência mundial, como a Organização das Nações Unidas - ONU e a Associação Internacional das Águas (*International Water Association* - IWA, que visam apoiar as nações e concessionárias de serviços de abastecimento de água a operacionalizar a captação e consumo racional e sustentável dos recursos hídricos.

3.1.1 Objetivos de Desenvolvimento do Milênio - ODM

A Organização das Nações Unidas - ONU, lidera agendas para o desenvolvimento mundial desde sua fundação, em 1940. Os Objetivos de Desenvolvimento do Milênio (ODM) foram adotados pelos 191 estados membros das Nações Unidas em setembro de 2000 que englobavam 8 ambiciosos objetivos e 22 metas que colocavam as pessoas e suas necessidades imediatas na linha de frente, remodelando os processos decisórios nos países desenvolvidos bem como os processos de desenvolvimento [13] [14] [15].

Figura 3.1: Objetivos do Milênio (Adaptado de http://www.odmbrasil.gov.br).

Os recursos naturais são vitais para garantir alimentação da população mundial bem como assegurar outras necessidades sociais, econômicas e ambientais, de modo que se torna imperativo sua exploração sustentável. As mudanças climáticas, conflitos sobre acesso aos recursos naturais e especialmente a crescente escassez de água representam uma ameaça, não apenas ambiental, mas também à segurança alimentar da população [16].

Por este motivo, o sétimo objetivo de desenvolvimento do milênio (ODM 7) que visava assegurar a sustentabilidade ambiental, estabeleceu um conjunto de 4 metas a serem perseguidas pelas nações participantes [17] :

a. Integrar às políticas dos países, princípios de desenvolvimento sustentável e programas para reversão das perdas de recursos ambientais;

b. Reduzir a perda de biodiversidade avançando, até 2010, a uma taxa significativa de redução;

c. Reduzir pela metade, até 2015, a proporção da população sem acesso sustentável a água potável e saneamento básico;

d. Alcançar, até 2020, uma melhora significativa na vida de pelo menos 100 milhões de moradores de favelas.

Destaca-se, para os objetivos da presente pesquisa, a ODM 7.c, que visava a implementação de ações que possibilitassem, a longo prazo, a universalização do acesso à água potável.

Neste sentido, o *Millenium Development Goal* afirma que em 2015 91% da população mundial passou a ter acesso a fontes de água potável em melhores condições que tinham em 1990, onde o índice era de 76%. Este índice representa um número absoluto de 2,6 bilhões de pessoas que ganharam acesso a uma melhor água potável desde 1990. Destas, 1,9 bilhões têm acesso à água canalizada nas instalações, com 58% da população mundial desfrutando desse nível de serviço em 2015 [16].

Apesar de aparentemente a meta da ODM7 ter sido atingida, segundo o Programa das Nações Unidas para o Desenvolvimento - PNUD, aproximadamente 2.2 bilhões de pessoas ainda carecem de acesso a um sistema seguro de gestão de recursos hídricos, e 4.5 bilhões, a um sistema seguro de saneamento. O PNUD destaca, ainda, que 70% da água potável do mundo é consumida pela agricultura e cerca de 1.9 bilhão de pessoas vivem em regiões onde pode faltar água em um futuro próximo [18].

3.1.2 Objetivos de Desenvolvimento sustentável - ODS

Como exposto no início deste capítulo, os ODM foram definidos no ano 2000 contemplando oito objetivos a serem alcançados até o final de 2015. Significativos progressos foram realizados, tendo os oito ODM's como plano de fundo, e fizeram a diferença na vida de bilhões de pessoas ao redor do mundo ajudando a minimizar a pobreza mundial, mas muito ainda há para se fazer.

Em setembro de 2015, refletindo os desafios apresentados pela conferência da ONU conhecida como Rio+20 ocorrida em 2012, a Organização das Nações Unidas definiram os 17 Objetivos de Desenvolvimento Sustentável (ODS), uma continuação dos Objetivos de Desenvolvimento do Milênio, também conhecido como Agenda 2030 [19]. Os ODS provêm à sociedade mundial (pessoas, governos, organizações, etc) um conjunto de objetivos que direcionam ações, público ou privadas, visando alcançar uma meta comum a todos: o desenvolvimento sustentável de todas as nações [20].

Figura 3.2: Objetivos do Desenvolvimento Sustentável.

Na Rio+20, a água foi reconhecida como questão nuclear do desenvolvimento sustentável, a chave determinante para todos os aspectos do desenvolvimento social, econômico e ambiental, de forma que deveria ter posição central nos debates sobre erradicação da pobreza e desenvolvimento sustentável global. Este foi um dos motivadores para que o *Open Working Group* das Nações Unidas, grupo consultivo que liderou a definição dos Objetivos para o Desenvolvimento Sustentável (ODS), apresentasse como um dos 17 ODS a garantia para disponibilidade e gerenciamento sustentável de águas e saneamento para todos (ODS 6) com seis metas distintas [20] :

1. Até 2030, alcançar o acesso universal e equitativo a água potável e segura para todos;

2. Até 2030, alcançar o acesso a saneamento e higiene adequados e equitativos para todos, e acabar com a defecação a céu aberto, com especial atenção para as necessidades das mulheres e meninas e daqueles em situação de vulnerabilidade;

3. Até 2030, melhorar a qualidade da água, reduzindo a poluição, eliminando despejo e minimizando a liberação de produtos químicos e materiais perigosos, reduzindo à metade a proporção de águas residuais não tratadas e aumentando substancialmente a reciclagem e reutilização segura globalmente;

4. Até 2030, aumentar substancialmente a eficiência do uso da água em todos os setores e assegurar retiradas sustentáveis e o abastecimento de água doce para enfrentar a

escassez de água, e reduzir substancialmente o número de pessoas que sofrem com a escassez de água;

5. Até 2030, implementar a gestão integrada dos recursos hídricos em todos os níveis, inclusive via cooperação transfronteiriça, conforme apropriado;

6. Até 2020, proteger e restaurar ecossistemas relacionados com a água, incluindo montanhas, florestas, zonas úmidas, rios, aquíferos e lagos:

   a. Até 2030, ampliar a cooperação internacional e o apoio à capacitação para os países em desenvolvimento em atividades e programas relacionados à água e saneamento, incluindo a coleta de água, a dessalinização, a eficiência no uso da água, o tratamento de efluentes, a reciclagem e as tecnologias de reuso;

   b. Apoiar e fortalecer a participação das comunidades locais, para melhorar a gestão da água e do saneamento.

Como se percebe pelas metas definidas para o ODS 6 listadas acima, a Agenda 2030 da ONU relacionada às ações de uso sustentável de água foi construída com as lições aprendidas na execução dos Objetivos de Desenvolvimento do Milênio 7 e aborda muitas de suas deficiências. O ODS 6 expande o escopo original da ODM 7 para incluir, além da universalização ao acesso á água e saneamento, a segurança, equidade e sustentabilidade [20].

A meta 6.4 da ODS 6 desperta especial interesse para o escopo desta pesquisa, uma vez que visa aumentar a eficiência no uso de água e o abastecimento da população de forma sustentável. Em um sistema urbano de abastecimento de água, a identificação e combate às fraudes no consumo são ações primordiais para combater as perdas de água tratada [9] [21], contribuindo para seu uso racional e, consequentemente, ao uso eficiente das águas, indo ao encontro da meta 4 do Objetivo de Desenvolvimento Sustentável 6, proposto pela ONU.

Reduzir as perdas é um dos focos na estratégia de assegurar o uso eficiente de água. Cole *at al* [22] compartilham deste entendimento ao propor, em recente trabalho, sugestão de indicadores adicionais para o ODS 6.4, entre eles, o monitoramento da perda de água nos sistemas de distribuição de água tratada.

Cole [22] destaca que alcançar os objetivos da ODS 6 é basilar para assegurar o atendimento aos demais objetivos de desenvolvimento sustentável, uma vez que a água é direito fundamental do ser humano, indispensável para o crescimento econômico (ODS 8), para a segurança alimentar (ODS 2) e para saúde da população (ODS 3). Portanto, ao propor medidas para alcançar a meta 6.4, indiretamente estamos contribuindo em muitas das demais 168 metas da Agenda 2030 da ONU.

3.1.3 Diretrizes internacionais para combate às perdas de água

O *International Water Association - IWA* é uma organização internacional que compõe a maior rede de profissionais que desenvolvem pesquisas e projetos focados em soluções de gerenciamento de água, organizam eventos de classe mundial que trazem as mais recentes ciências, tecnologias e melhores práticas para o setor de águas em geral. O IWA trabalha para colocar a água na agenda política global e influenciar as melhores práticas de regulamentação e formulação de políticas.

Segundo o IWA, a maioria dos sistemas de abastecimento de água ao redor do mundo experimenta alto nível de perda de água [21]. Confirmando esta afirmação do IWA, em fevereiro de 2018 o Sistema Nacional de Informações sobre o Saneamento - SNIS informou que em 2016 o índice médio de perdas de água no Brasil foi de 37% e no Distrito Federal de 35,21% [8].

Ao sugerir as melhores práticas na gestão de águas, o IWA elaborou uma abordagem padrão para calculo do balanço hídrico (Figura 3.3), que possibilita identificar as perdas nos sistemas de abastecimento de água.

<table>
<tr><td rowspan="10">System Input Volume</td><td rowspan="3">Authorised Consumption</td><td rowspan="2">Billed Authorised Consumption</td><td>Billed Metered Consumption (including water exported)</td><td rowspan="2">Revenue Water</td></tr>
<tr><td>Billed Unmetered Consumption</td></tr>
<tr><td rowspan="2">Unbilled Authorised Consumption</td><td>Unbilled Metered Consumption</td><td rowspan="8">Non-Revenue Water (NRW)</td></tr>
<tr><td>Unbilled Unmetered Consumption</td></tr>
<tr><td rowspan="5">Water Losses</td><td rowspan="2">Apparent Losses</td><td>Unauthorised Consumption</td></tr>
<tr><td>Metering Inaccuracies</td></tr>
<tr><td rowspan="3">Real Losses</td><td>Leakage on Transmission and/or Distribution Mains</td></tr>
<tr><td>Leakage and Overflows at Utility's Storage Tanks</td></tr>
<tr><td>Leakage on Service Connections up to point of Customer metering</td></tr>
</table>

Figura 3.3: Abordagem Padrão para Balanço Hídrico (Fonte: [21]).

Esta abordagem proposta pelo IWA tem sido adotada nos cinco continentes, por diversas agências reguladoras e companhias de abastecimento de água ao redor do mundo, tais como: Austrália, Alemanha, África do Sul, Brazil, Canada, Estados Unidos, dentre outros. Mesmo as companhias que já possuiam padrões de cálculo de perdas de água bem definido, tem adaptado seus métodos à abordagem sugerida pelo IWA [21].

Nesta esta abordagem, o IWA estabelece as seguintes definições:

- *System Input Volume:* Refere-se ao volume anual de entrada no sistema de abastecimento de água;

- *Autorised Consumption:* É o volume anual da água que foi efetivamente consumida, de forma autorizada, por clientes cadastrados;
- *Water Losses:* É a diferença entre o volume de entrada (*System Input*) e o consumo autorizado (*Authorised Consumption*)

Como ilustra a Figura 3.3, as perdas de água podem ser reais ou aparentes, sendo a primeira motivada por vazamentos nas tubulações dos sistemas de abastecimento e a segunda pelo consumo não faturado da água distribuída.

Perdas aparentes são motivadas, basicamente, por quatro fatores: imprecisão nos medidores de consumo, erro nos dados de medição, consumo não autorizado (roubo, ligações irregulares, fraudes nos medidores, etc) e erro de faturamento. Além de impactar significativamente na receita das companhia de abastecimento de água, este tipo de perda distorce dados de consumo, relevantes para análises e tomadas de decisões gerenciais bem como para estudos de engenharia, impactando na eficiência operacional [10].

As perdas aparentes são mais acentuadas em companhias de abastecimento de países em desenvolvimento. Estima-se que estes países enfrentem um prejuízo de cerca de 3 bilhões de dólares por ano com perdas aparentes [10].

O cálculo de perdas aparentes são baseados em testes de amostragem estruturada ou por estimativa apoiada por robustos procedimentos definidos em propósito de auditorias, que resultam em um percentual de perdas aparentes considerando o volume de entrada. No entanto, a IWA recomenda que cada companhia de água defina métodos próprios para realizar avaliação precisa das perdas aparentes dos sistemas de distribuição de água que operam [21].

Não obstante, a maior parte das pesquisas referentes a perdas de água focaram nas perdas reais, ou seja, em problemas de vazamentos. Em trabalho recente, Mutikanga *et al* [10] afirmam que procedimentos ou guias para avaliação de perdas aparentes são inexistentes e que na ausência de dados adequados e metodologias próprias, em geral adota-se valores padrão para perdas aparentes nos cálculos de balanço hídrico.

Por este motivo, para quantificar com precisão as perdas de água aparentes, imperioso se faz a aplicação de um método sistematizado que possibilitará às empresas prestadoras de serviços de abastecimento minimizar seus riscos de perdas operacionais e financeiras.

3.1.4 Ações adotadas pela Caesb para combate as perdas de água

A Companhia de Saneamento Ambiental do Distrito Federal - Caesb tem intensificado os investimentos na prevenção e combate as perdas de água.

Em 2017 foram investidos R$ 170 milhões em ações preventivas e corretivas de fatores que contribuem para as perdas reais. [1]

Ao longo de 2018, foram executadas obras de setorização da rede, telemetria do sistema distribuidor, mapeamento e monitoramento de zonas de pressão, estudos de revitalização do sistema para substituição de redes antigas, entre outras ações de engenharia que contribuem para redução das perdas reais.

No que tange as perdas aparentes, cujos principais fatores são i) submedição do consumo, ii) consumo não autorizado (ligações clandestinas) e iii) consumo não medido e\ou não faturado (erros nos dados ou fraudes de consumo), a companhia tem realizado investimentos para atacar os dois primeiros itens.

Para minimizar as perdas por submedição, em 2016 foram substituídos pouco mais de 53 mil hidrômetros em todo o Distrito Federal, o que resultou em um retorno financeiro na ordem de R$ 100 mil em 2017.

Para atacar os consumos não autorizados, foram realizadas iniciativas utilizando Sistemas de Informações Geográficas - SIG (ou como é mais conhecido, GIS *Geographic Information Systems)* para identificar possíveis edificações com consumo irregular de água, adotando a técnica de fotointerpretação onde analisou-se a mancha urbana de todo o Distrito Federal sobreposta por uma camada de dados espaciais contemplando as redes de distribuição de água e as ligações regulares. As edificações visíveis na Ortofoto em que não constasse uma ligação regular eram marcadas com um ponto vermelho como possível ligação irregular (Figura 3.4), seu consumo médio estimado pelo padrão de consumo da vizinhança e as perdas calculadas pelo somatório do consumo estimado destas possíveis ligações irregulares.

Figura 3.4: Identificação de consumo não autorizado.

[1] fonte: https://www.agenciabrasilia.df.gov.br/2018/01/29/caesb-reduz-perdas-de-agua-em-2017-e-reforca-combate-aos-desvios/

Este trabalho resultou na identificação de mais de 40 mil possíveis ligações irregulares, com um prejuízo operacional estimado em 8,7 milhões de m3/ano e um prejuízo financeiro na ordem de R$ 43 milhões de reais/ano [2].

O trabalho realizado em menos de um mês na região do Sol Nascente, cidade satélite de Ceilândia no Distrito Federal possibilitou regularizar pouco mais de 4 mil ligações, o que resultou em um retorno financeiro na ordem de R$ 3,7 milhões por ano.

Ações de regularização destas ligações identificadas como possível consumo irregular possibilitou a redução das perdas aparentes no questito "consumo não autorizado".

No entanto, o item iii, "consumo não medido e\ou não faturado"continua sendo uma variável desconhecida no processo de prevenção e combate a perdas aparentes.

Visando preencher esta lacuna, o presente trabalho pretende elaborar um modelo para identificação de riscos ao faturamento causados por fraudes no consumo que, somada as outras iniciativas já adotadas pela Caesb, vão ao encontro dos objetivos da ODS 6.4 da Organização das Nações Unidas (Capítulo 3.1.2) que é o uso eficiente dos recursos hídricos, ao mesmo tempo que contribui para a redução dos riscos de perdas financeiras e operacionais percebidos pela empresa.

3.2 *Cases* reais de identificação de riscos ao faturamento

Nesta sessão são apresentados *cases* reais da aplicação prática de modelos que visam identificar riscos ao faturamento por meio da detecção de fraudes nos mais diversos setores da economia.

A intenção desta sessão é explorar as principais técnicas, ferramentas e modelos aplicados para resolver problemas de faturamento ocasionado por fraudes, em qualquer que seja o campo de aplicação, e avaliar possível adequação ou combinação destas para proposição do modelo a ser aplicado no estudo de caso deste trabalho.

3.2.1 Setor Elétrico

A maioria dos casos de fraudes no consumo de energia elétrica são combatidos por meio de inspeção direta realizada por técnicos das companhias de distribuição de energia elétrica de modo que, visando minimizar os custos e conferir maior produtividade, as inspeções precisam ser realizadas em locais que efetivamente apresentam problemas de consumo irregular [23].

[2]Fonte: https://www.agenciabrasilia.df.gov.br/2017/09/24/projeto-atlas-indica-prejuizo-de-r-43-milhoes-com-ligacoes-clandestinas-de-agua/

Neste sentido, para garantir maior assertividade nas campanhas de combate as fraudes no consumo de energia elétrica de uma empresa de distribuição da Espanha, Puig *et al* [23] elaboraram um modelo de detecção de fraudes baseado em aprendizagem supervisionada para classificação de possíveis clientes com condutas fraudulentas.

Uma característica interessante do modelo proposto por Puig é a retroalimentação do modelo (Figura 3.5) que permite que novas campanhas sejam cada vez mais precisa considerando os resultados de campanhas anteriores.

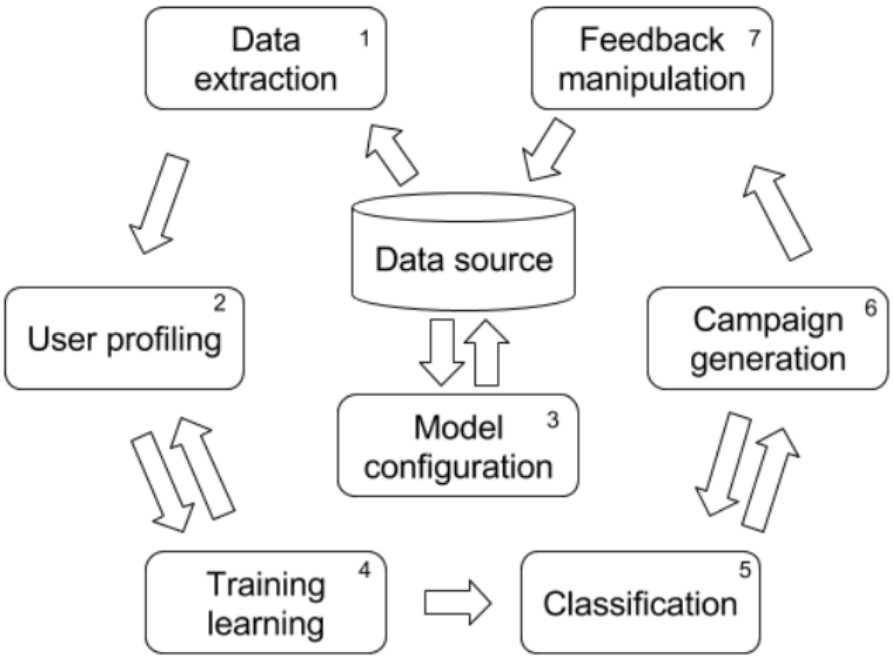


Figura 3.5: Modelo de detecção de fraudes de Puig (Fonte: [23]).

A primeira campanha de identificação e combate a fraudes utilizando este modelo foi realizada em uma população de até 100 mil habitantes e resultou em um aumento de 5 vezes na precisão da indicação de locais com indícios de fraudes.

Complementarmente, Ahmad *et al* [24] realizaram uma revisão da literatura sobre aplicação de técnicas de mineração de dados para detecção de fraudes no consumo de energia elétrica e afirmam que são extremamente úteis na predição e detecção de fraudes.

3.2.2 Setor Financeiro

Fraudes financeiras não são exclusividades de operações que envolvam intercâmbio de valores ou movimentação de moedas. As fraudes em declarações financeiras, aquelas que ocorrem em documentos que refletem a saúde de uma empresa, tem crescido nos últimos

anos causando graves impactos ao desenvolvimento sustentável de empresas e do próprio mercado financeiro, de modo que a definição de um modelo de alerta para detecção de atividades fraudulentas nos relatórios financeiros de grandes corporações tem sido objeto de estudos e discussões no meio acadêmico [25].

Jan [25] em recente estudo que contemplava 160 companhias presentes no mercado financeiro de Taiwan afirma que as fraudes de declarações financeiras são problemas típicos de classificação. O autor adotou diversas técnicas de mineração de dados para elaboração de um modelo de detecção de fraudes que resultaram na classificação de fraudes com acurácia de 90.83%.

Ngai *et al* [26] também realizaram uma revisão da literatura sobre aplicação de mineração de dados nos diversos setores de detecção de fraudes financeiras, tais como fraudes a seguradoras, fraudes bancárias e fraudes de cartão de crédito e estudos realizados por pesquisadores da China e Estados Unidos resultaram em uma proposta de um *framework* a ser aplicado em detecção de fraudes financeiras utilizando mineração de dados (Figura 3.6).

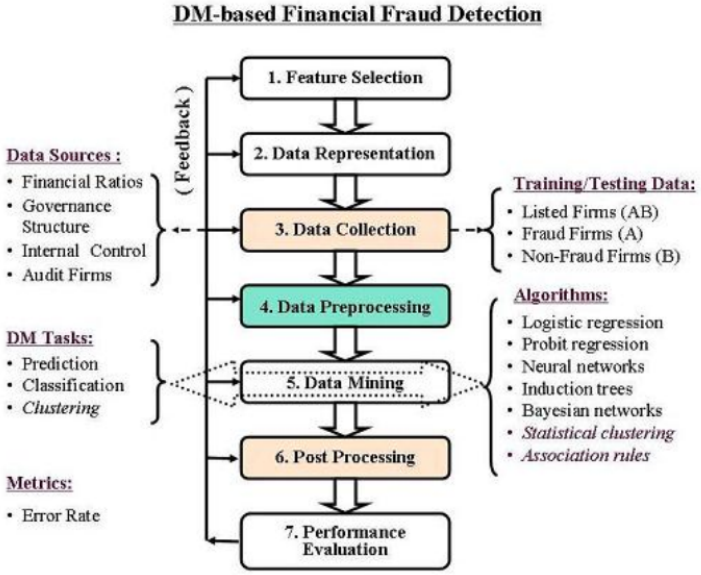


Figura 3.6: *Framework* para detecção de fraudes (Fonte: [27]).

Este *framework* para detecção de fraudes mantem estreita semelhança com os processos de mineração de dados tradicionais e internacionalmente consolidados como KDD e Crisp-DM, com etapas bem definidas para seleção, preparação dos dados, mineração e avaliação dos resultados.

3.2.3 Setor de Crédito Individual

Fraudes em cartão de crédito são caracterizadas pelo uso indevido, sem conhecimento ou autorização do proprietário do cartão [28]. Milhões de operações com cartão de crédito são realizadas diariamente, gerando um enorme volume de dados de modo que aplicação de técnicas de *data mining* se torna imperativa para detecção de fraudes.

Neste contexto, Carneiro [29] propôs uma abordagem para estimar o *score* de suspeição de fraudes de operações com cartão de crédito em um empreendimento online de artigos de luxo aplicando as técnicas de mineração de dados *random forests, logistic regression* e *support vector machine* com bons resultados, sendo que o primeiro apresentou melhor performance para tratar com grande volume de dados e ser de fácil implementação.

Kho e Vea [30] avaliaram técnicas de mineração de dados usando árvore de decisão para classificar operações com cartão de crédito visando identificar operações fraudulentas e afirmaram que *Random Tree* e *J48* apresentaram melhores resultados, com acurácia de 94,32% e 93,50% respectivamente.

Randhawa *et al* [28] propuseram um modelo híbrido para detecção de fraudes em cartões de crédito composto por múltiplos métodos de mineração de dados contemplando Naive Bayes, Máquina de Vetores de Suporte, AdaBoost e Majority voting.

3.2.4 Setor de Seguro Saúde

O grande volume de dinheiro envolvido nas atividades de seguro saúde tornam este setor alvo de fraudes. As fraudes em seguros saúde são caracterizadas pela realização de procedimentos médicos desnecessários ou reembolso de serviços não prestados, onerando as seguradoras e beneficiando o paciente ou entidade prestadora de serviço.

Verma *et al* [31] afirmam que recentes avanços nas políticas de seguros saúde requerem adoção de sistemas sofisticados para detecção fraudes, levando invariavelmente a utilização de métodos de mineração de dados. Neste sentido, os autores propuseram uma abordagem de detecção de fraudes que contempla técnicas de associação, clusterização e detecção de *outliers*.

Hillerman *et al* [32] utilizaram métodos de clusterização e algorítimo *k-means* para agrupar e identificar operações suspeitas em um provedor de seguro saúde do Brasil e o método multicritério AHP para priorizar as operações a serem investigadas.

3.2.5 Setor de Saneamento

Fraudes tem potencial impacto negativo no faturamento de diversas atividades econômicas de modo que métodos e tecnologias de detecção de fraudes tem despertado interesse de

empresas nos mais diversos segmentos, como telecomunicação, instituições de crédito, seguros, entre outros [11].

No que tange ao consumo de água urbana, as fraudes são definidas como intervenções voluntárias nos medidores de consumo (hidrômetros) objetivando mascarar a medição de forma que apenas parte da água efetivamente consumida seja computada [7]. A adulteração de hidrômetros é uma manipulação dos medidores de consumo de água que causam prejuízos as empresas de abastecimento de água, uma vez que os serviços prestados não são efetivamente pagos [33].

Métodos de detecção de fraudes permitem assegurar maior assertividade nos planejamentos operacionais, pois possibilitam direcionar recursos e esforços em locais com maior probabilidade de irregularidade no consumo de água. No entanto, detecção de fraudes em consumo de água é uma tarefa complexa, uma vez que diversas variáveis, tais como variabilidade no padrão de consumo, sazonalidade, alteração na quantidade de moradores, entre outros fatores, devem ser considerados [7].

Passini *et al* [11] aplicaram a mineração de dados na Companhia de Saneamento de Campinas - SANASA, onde estima-se que dos 26% de perdas de água ocorrida no ano 2000, 5% tenha sido causado por fraudes no consumo, visando identificar perfil de clientes que apresentam características que apontam para indícios de condutas fraudulentas no consumo de água.

Os pesquisadores criaram três modelos de mineração: dois agrupamentos neurais e um classificação por árvore de decisão. Nos agrupamentos, buscava-se identificar um perfil de fraude considerando o padrão de consumo e categoria do cliente, de modo a possibilitar identificar os consumidores que se encaixam neste perfil para realizar eventuais vistorias em campo. O modelo baseado em classificação, por sua vez, busca predizer qual tipo de fraude os possíveis consumidores fraudadores cometeriam [11].

Humaid *et al* [27] elaboraram um modelo de predição de fraude de consumo de água em 67 regiões da cidade de Gaza. O modelo utilizou a técnica de mineração de dados *rule induction* e destacou regiões com potencial de fraudes no consumo. Segundo os autores, os pontos mais relevantes de seu trabalho foi identificar as perdas financeiras causadas pelas fraudes de consumo além de detectar localidades com perdas aparentes mais significativas.

Fetterman *et al* [7] realizaram estudo de detecção de fraudes no consumo de água no município de Jequié, Bahia, Brasil e propuseram uma sistemática composta por 5 etapas que apoiam na detecção das irregularidades (Figura 3.7).

Necessidade	Etapa	Descritivo
Identificar o comportamento dos dados e suas características	1. Análise Descritiva dos dados	Visa compreender e interpretar características, tais como medidas de tendência central e dispersão, além de verificar o comportamento da serie temporal.
Localizar o método mais adequado para identificar fraudes no caso analisado	2. Seleção do método para detecção de fraudes	Procura-se verificar, a partir das suposições de aplicação dos três métodos, o mais adequado para detecção de fraudes no caso analisado.
Adequar o comportamento dos dados selecionados para viabilizar a aplicação do método escolhido	3. Tratamento dos dados	Verificar necessidade de algum tipo de transformação nos dados ou redução de sua variabilidade para o melhor desempenho da técnica selecionada.
Realizar aplicação do método de detecção de fraudes escolhido	4. Aplicação do método selecionado	Aplicar o método selecionado seguindo as recomendações propostas pela bibliografia.
Analisar a capacidade em detectar fraudes no caso estudado	5. Análise dos resultados	Identificar a taxa de deacerto na identificação das economias com fraude e a taxa de indicação de fraude em casos em que não foi detectado (falso positivo).

Figura 3.7: Proposta de sistemática para detecção de fraudes (Fonte: [7]).

Neste estudo, os autores utilizaram o método de detecção de *outliers* e a técnica *Z-score* para identificar, por meio da análise de média e desvio padrão do consumo, variações que possam indicar fraudes no consumo de água. O resultado obtido pelos autores possibilitou melhor direcionamento dos esforços de inspeção de fraudes no sistema de abastecimento.

Candelieri [34] aplicou técnicas de *machine learning*, uma confluência da mineração de dados, na cidade de Milão, Itália, para caracterizar e prever a demanda de consumo de água. Este trabalho possibilitou identificar anomalias no consumo, como possíveis fraudes.

Monedero *et al* [33] propuseram uma metodologia de detecção de baixo consumo anormal ou redução de consumo suspeita para os clientes de uma empresa de abastecimento de água de Seville, Espanha. A metodologia utiliza um conjunto de algoritmos de mineração de dados.

Percebe-se, portanto, que a revisão da literatura sobre aplicações de detecção de fraudes no consumo de água urbana, apesar de escassas, demonstram uma tendência na adoção de modelos que utilizam métodos de mineração de dados como ferramenta de apoio.

3.3 Mineração de Dados

A sessão anterior apresentou exemplos de aplicações práticas para detecção de fraudes nos mais diversos setores econômicos. Percebe-se nos *cases* apresentados que os modelos de detecção de fraudes, independente do ramo de aplicação, utilizaram diferentes técnicas de mineração de dados para identificação de indícios de ações ou comportamentos fraudulentos com excelentes resultados.

Apesar da diversidade de técnicas aplicadas, o processo para mineração de dados em todos os *cases* pesquisados são similares, salvo pequenas adaptações para se adaptar à realidade fática de cada *case*.

É razoável entender, portanto, que a mineração de dados pode ser aplicada no presente estudo de caso com a mesma eficiência que tem-se notado em outras aplicações.

Portanto, mister se faz aprofundar o entendimento nos conceitos e processos da mineração de dados para elaboração de um modelo que atenda os objetivos deste trabalho.

Neste sentido, esta sessão apresenta uma sínteses objetivas da revisão da literatura onde buscou-se identificar os principais processos e conceitos de mineração de dados que subsidiarão a elaboração modelo resultante desta pesquisa.

3.3.1 Processos de Mineração de Dados

Há algum tempo, as organizações vem produzindo uma enorme quantidade de dados, tais como dados de clientes, histórico de transações, registros de vendas, etc. Estes dados são extremamente valiosos para as organizações [35].

A mineração de dados permite a indexificação de padrões potencialmente úteis e de fácil entendimento [36] [11], bem como a descoberta de interessantes e desconhecidas tendências, padrões e relacionamentos nos conjuntos de dados que, de outra forma, permaneceriam desconhecidos. Em outras palavras, ela revela informações ocultas sobre um grande volume de dados [37].

A detecção de fraudes por meio da mineração de dados e análise de comportamento do consumidor permite a geração de novos conhecimentos atraves de um modelo do mundo real que descreve padrões e relacionamentos que podem ser utilizados para realizar predições [11].

Em última análise, a mineração de dados é a extração de conhecimento que utiliza métodos computacionais e intervenção humana em diversas etapas do processo, desde a

coleta e preparação de dados e, principalmente, na análise dos resultados para tomada de decisão [35].

O processo de mineração de dados é cíclico e iterativo, que se inicia com uma hipótese e utiliza-se dados para refutá-la ou confirmá-la. A hipótese é redefinida e o processo continua até que se atinja uma resposta satisfatória [35].

Azevedo e Santos [38] demonstram em seu artigo *KDD, SEMMA and CRISP-DM: A parallel overview* que os processos ou métodos analisados tem por objetivo extrair conhecimento útil para o negocio, a partir da mineração de bases de dados por meio de atividades que envolvam tratamento de inconsistências, identificação de padrões e extração de conhecimento. O que difere entre os processos ou metodos estudados é a forma e etapas necessárias para chegar a este fim, conforme descrito nas subseções abaixo.

KDD - Knowledge Discovery in Database

Uma sistematização do processo de mineração foi proposta por Fayyad [39] em 1996 e se sedimentou como metodologia basilar para descoberta de conhecimento em banco de dados, ou como é mundialmente conhecido: *Knowledge discovery in Database - KDD* (Figura 3.8).

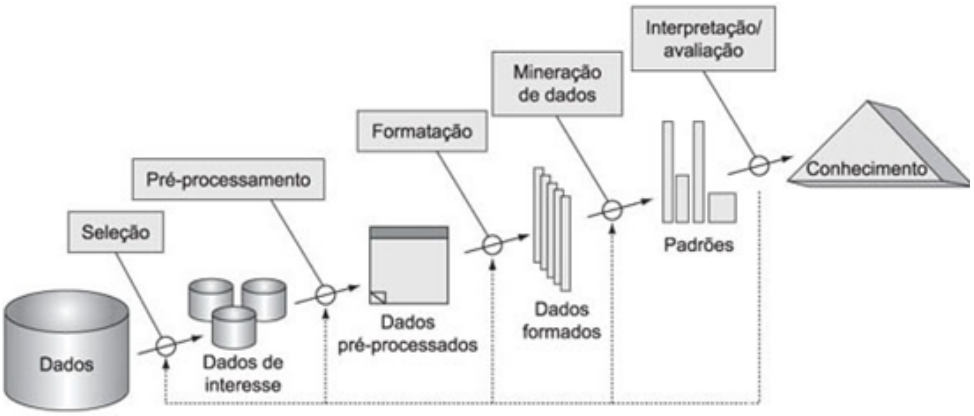


Figura 3.8: Processo KDD (Fonte: [39]).

O processo de KDD proposto por Fayyad inicia-se pela seleção dos dados de interesse a partir de um grande conjunto de dados de entrada, na sequência, os dados selecionados devem ser pré-processados que consiste na limpeza (eliminação de dados incompletos), transformação ou enriquecimento (ajustes nos atributos dos dados), a próxima etapa consiste na mineração em sí que resulta na extração de padrões dos dados selecionados e, por fim, a interpretação da mineração de dados, última etapa do processo onde gera-se conhecimento.

O KDD prevê cinco estágios para mineração de dados:

1. **Seleção:** extração de amostras significativas dos dados de uma base de dados em que serão executadas as próximas atividades de mineração
2. **Pre-processamento**: etapa em que atividades para consistência dos dados são executadas
3. **Transformação**: nesta etapa, os dados já consistidos são "transformados", usando métodos de redução dimensional.
4. **Data mining**: nesta etapa, padrões são identificados, considerando os objetivos do negócio.
5. **Interpretação**/Avaliação: Etapa final do processo KDD, que consiste na interpretação dos padrões minerados na etapa anterior.

SEMMA - Sample, Explore, Modify, Model, Assess

SEMMA, acrônimo de *SAMPLE, EXPLORE, MODIFY, MODEL, ASSESS*, é um processo criado pela SAS Institute para guiar as atividades de mineração de dados (Figura 3.9).

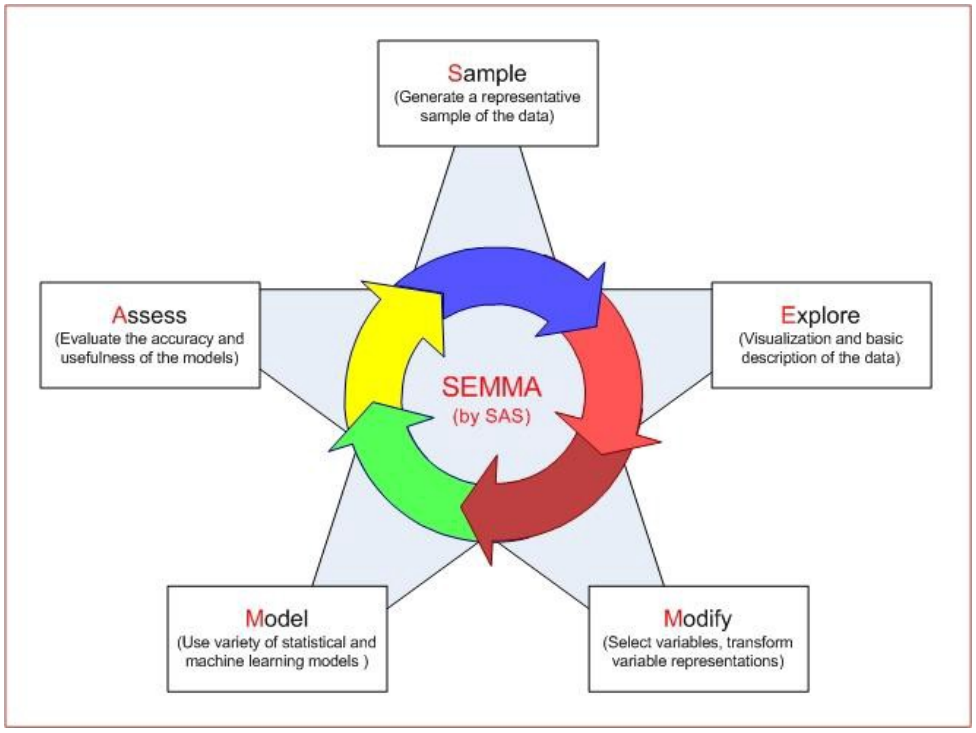


Figura 3.9: Processo SEMMA (Fonte: [38]).

Assim como o KDD, a SEMMA orienta realizar a mineração de dados pela execução de 5 etapas:

1. **Exemplo**: Refere-se a extração de uma amostra significativa dos dados
2. **Exploração**: Nesta etapa, busca-se explorar os dados extraídos como amostras no intuito de entender sua organização e identificar possíveis anomalias.
3. **Modificação**: Nesta etapa, os dados são manipulados para criar, selecionar e transformar variáveis que serão adotadas no modelo de mineração.
4. **Modelo**: Nesta etapa, é gerado um modelo para extração de dados que prevejam os resultados desejados.
5. **Avaliação**: Esta etapa consiste em avaliar os dados considerando a utilidade e a confiabilidade dos resultados do processo de DM e o desempenho.

CRISP-DM - Cross Industry Standard Process for Data Mining

CRISP-DM é a abreviação de *CRoss Industry Standard Process for Data Mining*, elaborado pela CRISP-DM consortium para ser um *framework* de referência para mineração de dados independente da industria, podendo ser aplicado em qualquer área de negócio, como por exemplo comercio, financeiro, recursos humanos, produção agrícola, dentre outros (Figura 3.10).

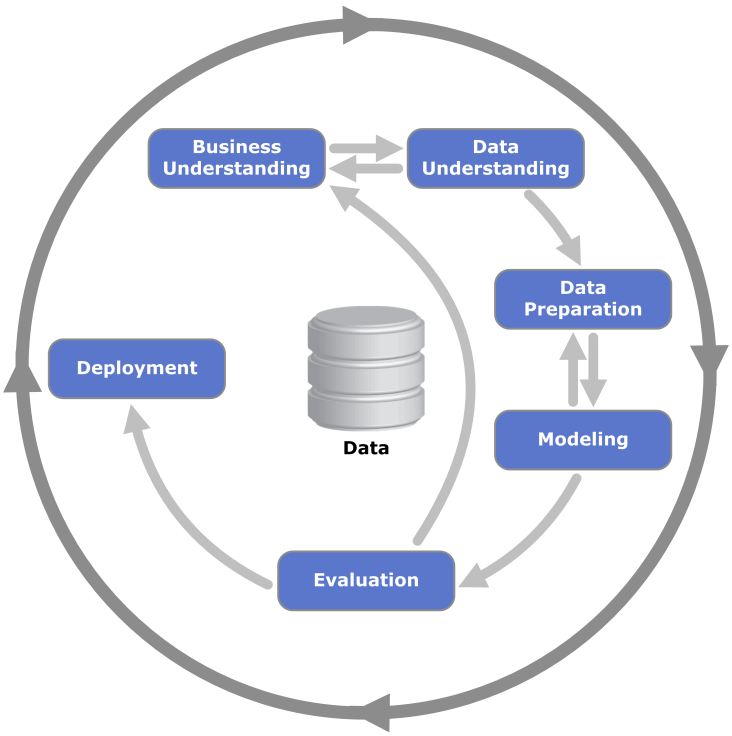


Figura 3.10: Processo CRISP-DM (Fonte: [38]).

Assim como os processos anteriores, o CRISP-DM prevê uma abordagem cíclica para a mineração de dados, inovando, se comparado aos processos anteriores, ao prever uma etapa de entendimento do negócio.

Ao total, o CRISP-DM prevê seis etapas para mineração de dados:

1. **Entendimento do Negócio**: Esta etapa está focada no entendimento dos objetivos e requisitos do negócio para uma extração de dados e conhecimento efetiva.

2. **Entendimento dos dados**: Nesta etapa, uma amostra dos dados é selecionada para entender sua estrutura e organização, identificar problemas de qualidade e consistência dos dados

3. **Preparação dos dados**: Nesta etapa, são desempenhadas atividades para construção de um conjunto de dados consolidados a partir dos dados analisados na etapa anterior.

4. **Modelagem**: Nesta etapa, várias técnicas de modelagem são aplicadas para otimizar os resultados.

5. **Avaliação**: Nesta etapa os resultados dos modelos construídos são analisados mais detalhadamente e reconstruídos para alcançar resultados mais adequados aos objetivos do negócio.

6. **Desenvolvimento**: Nesta etapa final, o conhecimento obtido pela utilização do modelo gerado deve ser tratado para melhor compreensão do usuário final, que fará uso deste conhecimento para tomada de decisão.

3.3.2 Métodos de Mineração de dados

A mineração de dados se divide em diferentes métodos, cada um com capacidade para resolver problemas específicos:

1. **Classificação** cria um modelo baseado em dados rotulados (classe) já conhecidos para analisar os fatores da classificação existente que considera características dos elementos de cada classe, para classificar novos elementos desconhecidos, de modo a particionar os dados em grupos distintos auxiliando na tomada de decisão ao identificar padrões relevantes para o negócio em um conjunto de dados [35] [26]. As técnicas de classificação mais comuns são: Redes neurais, Naive Bayes, Árvores de Decisão e Máquina de vetores de Suporte.

2. **Associação** busca encontrar conexões entre elementos, onde a partir de uma regra resulta-se em uma consequência com um determinado fator de confiabilidade. Um exemplo de associação é a declaração de que 90% dos clientes que compram pão e manteiga também compram leite. Comprar pão e manteiga é a regra, comprar leite é a consequência e os 90% é o fator de confiabilidade da associação. Os problemas de associação auxiliam, dentre outros, a fortalecer a venda de produtos, por exemplo, encontrar todas as regras em que a consequência seja "Compram leite"possibilita planejar o melhor local em que o leite deve ficar disponível para aumentar a possibilidade de venda deste produto ou analisar o impacto que a ausência de um produto possa causar, como por exemplo, se faltar pão, pode-se reduzir a venda de leite [35] [11].

3. **Clusterização** agrupa objetos em grupos (*clusters*) conceitualmente semelhantes, onde os elementos de cada grupo são similares entre sí e diferem de elementos de outros grupos [26]. Difere-se da classificação pois a clusterização não utiliza rótulos, ou classes, para realizar o agrupamento, mas tão somente características inerentes a cada elemento. As técnicas mais comuns dessa classe de mineração de dados são: *K-nearest neighbor*, Naive Bayes, Mapa de alto organização.

4. **Predição** Estima valores numérico e ordens futuras a partir de um padrão identificado em um conjunto de dados [26]. As técnicas mais comuns são Redes Neurais e Modelo logístico de predição.

5. **Detecção de *outliers*** é aplicada para destacar elementos que se diferem de outros elementos em um mesmo conjunto de dados considerando o padrão definido pelas suas características [26]. A técnica mais comum desta classe é a *discounting learning algorithm*.

6. **Regressão** é um método estatístico utilizado para revelar a relação entre uma ou mais variável independente e uma variável dependente. As principais técnicas deste método são regressão logística ou regressão linear [26].

Os métodos de mineração de dados acima elencados se dividem em duas abordagens: supervisionados e não supervisionados [29].

A abordagem supervisionada requer um treinamento do algoritmo de mineração, ou seja, ensina-se para o algorítimo determinadas características do conjunto de dados para definição de um padrão esperado. Classificação, Associação e Predição são métodos de mineração de dados que utilizam esta abordagem.

A abordagem não supervisionada, por sua vez, não exige este treinamento, eles definem os padrões por meio da observação de características semelhantes nos dados analisados.

Esta abordagem é aplicada nos métodos de mineração de dados Clusterização, detecção de *Outliers* e regressão.

3.4 Gestão de Riscos

O objetivo primário da presente pesquisa, como mencionado em 1.2.1, é a identificação de riscos ao faturamento da Companhia de Saneamento Ambiental do Distrito Federal. Para conferir maior efetividade e aumentar a chance de sucesso de alcançar o objetivo deste trabalho, não se pode ignorar as melhores práticas preconizadas pela ISO 31.000, reconhecida mundialmente como referencia nos processos de Gestão de Riscos.

Portanto, visando sustentar o modelo proposto nesta pesquisa com um embasamento teórico sólido, este capítulo apresentará uma visão geral sobre o framework de gestão de riscos sugerido pela referida ISO.

3.4.1 ISO 31.000 - Princípios e Diretrizes na Gestão de Riscos

A ISO 31.000 foi elaborada por um comitê internacional composto por grupo de trabalho com representantes de 28 países com a contribuição de milhares de especialistas em gestão de riscos ao redor do mundo, que tinham por objetivo padronizar vocabulários e conceitos, definir critérios de desempenho e criar um processo abrangente de gestão de riscos aplicável a toda e qualquer atividade [40].

Segundo a ISO 31.000 [41], os processos de avaliação de riscos, uma vez aplicado a qualquer nível da organização, possibilitar obter dados e informações pertinentes e confiáveis para tomadas de decisões que englobem:

- Decidir a viabilidade de realização de determinada atividade;
- Como maximizar as oportunidades;
- Se e como tratar riscos identificados.

A ISO 31.000 sugere, ainda, que os processos de avaliação de riscos sejam implementados em conjunto com outras atividades, como a comunicação e consulta às partes interessadas e o monitoramento e análise crítica das ações adotadas. A Figura 3.11 ilustra o inter-relacionamento dos processos que compõe a estrutura ideal para Gestão de Riscos.

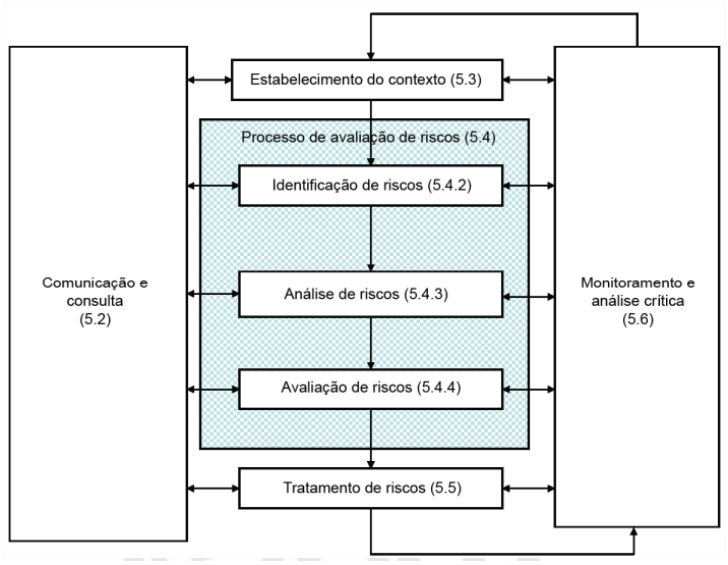


Figura 3.11: Processo de Gestão de Riscos (Fonte: [41]).

O processo proposto pela ISO 31.000 considera que na gestão de riscos, não basta identificar, analisar e avaliar seus impactos para os objetivos da organização, mais do que isto, deve ser considerado no processo de gerenciamento dos riscos, a comunicação e consulta às partes interessadas, além de análise e monitoramento para assegurar que nenhum tratamento de risco adicional seja necessário.

Comunicação e Consulta

Durante todas as fases do processo de gestão de riscos, as partes interessadas devem ser constantemente consultadas e comunicadas para garantir o alinhamento das expectativas, riscos e eventuais tratamentos a serem adotados.

A ISO 31.000 sugere que a comunicação e consulta sejam desenvolvidas com base em um plano de comunicação abordando questões relacionadas aos riscos, como suas causas, consequências e medidas tomadas para mitigar ou eliminar os riscos identificados.

A participação efetiva de uma equipe consultiva composta pelos principais interessados da organização contribuem para o processo de gestão de riscos pois:

- Auxilia a estabelecer um contexto mais assertivo;
- Assegura que os interesses e objetivos das partes interessadas sejam considerados;
- Assegura uma adequada identificação de riscos;
- Possibilita reunir diferentes áreas da organização para análise dos riscos;

- Assegura que diferentes pontos de vistas sejam considerados nas definições de critérios de identificação e avaliação dos riscos

Estabelecimento do Contexto

Não obstante os benefícios advindos de um processo bem definido para avaliação de riscos, é necessário estabelecer o contexto em que essa avaliação ocorrerá, garantindo assim, focar esforços apenas nos riscos inerentes ao contexto estabelecido.

Segundo a ISO 31.000, ao estabelecer o contexto, a organização articula seus objetivos e define parâmetros internos e externos que devem ser considerados para gerenciar riscos.

O contexto deve ser estabelecido em duas dimensões:

- **Contexto externo** é o ambiente externo à organização, envolvendo questões ambientais, culturais, sociais, políticos, legais ou regulatórios, tecnológicos, competitivo, dentre outros.
- **Contexto interno** é algo dentro da própria organização que pode influenciar a maneira como gerencia os riscos, como estrutura organizacional, força de trabalho, estratégias e diretrizes, etc.

Ainda no âmbito do estabelecimento do contexto, a ISO 31.000 sugere que sejam definidos critérios a serem utilizados para avaliar a significância dos riscos. Estes critérios devem considerar os valores, objetivos e recursos da organização, em especial aqueles elencados nos contextos internos e externos, uma vez que alguns critérios podem ser impostos por fatores externos como requisitos legais ou regulatórios.

Identificação, análise e avaliação de riscos

A ISO 31.000 engloba no processo de avaliação de riscos as atividades de identificação, análise e avaliação de riscos.

Identificação de Riscos, nesta etapa busca-se determinar o conjunto de eventos e fatores, externos ou internos, fontes de riscos, áreas de impacto, causa e consequências que podem impactar nos objetivos da organização. A identificação de riscos precede a avaliação dos riscos e produz uma lista compreensiva dos riscos a serem gerenciados.

Além da identificação dos riscos, suas fontes e causas, é necessário considerar possíveis causas e cenários que demonstrem as consequências da materialização dos riscos identificados.

O Instituto Brasileiro de Governança Corporativa afirma que o processo de identificação de riscos pode resultar em oportunidades, de modo que é imprescindível a participação de pessoas qualificadas e com visão holística da organização [42].

Análise de Riscos, esta etapa envolve desenvolver uma compreensão profunda dos riscos, sendo entrada para a atividade de avaliação e para decisão sobre estratégias e métodos mais adequados para o tratamento dos riscos identificados.

Durante a análise de riscos, as causas e fontes são estudadas, bem como suas consequências positivas e negativas e a probabilidade dessas ocorrências se materializarem. Portanto, considerando que os riscos são avaliados pela consequencia e impacto, é necessário que os fatores que afetam as consequências e a probabilidade de ocorrência sejam identificados e monitorados.

Avaliação de Riscos, tem por finalidade auxiliar na tomada de decisão com base na identificação e análise dos riscos, sobre quais riscos devem ser tratados, bem como prioridade para implementação do tratamento. Esta etapa envolve comparar o risco encontrado durante a etapa de identificação e análise com os critérios definidos durante o estabelecimento do contexto. Essa comparação subsidiará a decisão da necessidade e forma de tratamento do risco.

A ISO 31.000 sugere que as decisões para tratamento dos riscos considere a tolerância aos riscos assumidas pelas partes interessadas, de forma que a decisão pode ser simplesmente aprofundar mais nas atividades de análise ou simplesmente manter os controles existentes.

Tratamento de riscos

O tratamento de riscos envolve a seleção de uma ou mais opções para modificar os efeitos e impactos da ocorrência de um risco identificado e a implementação das ações necessárias.

A ISO 31.000 alerta que o tratamento dos riscos é um processo cíclico, onde deve-se:

- avaliar os tratamentos já realizados;
- decidir se os níveis residuais do risco são tolerados;
- definir e implementar novo tratamento de risco, caso o resíduo não seja tolerável; e
- avaliar a eficácia do novo tratamento.

As opções de tratamento não são necessariamente exclusivas. Mais de uma opção podem ser implementadas, tais como:

- remover a fonte do risco;
- alterar a probabilidade do risco;
- alterar a consequencia do risco;
- compartilhar o risco com outras partes;

- evitar o risco descontinuando ou não iniciando uma atividade.

Selecionar a opção mais adequada para o tratamento dos riscos envolve equilibrar os custos e esforços para sua implementação com os benefícios decorrentes, de modo que deve-se buscar a razoabilidade no tratamento dos riscos, em especial no critério econômico.

Monitoramento e análise crítica

A ISO 31.000 sugere que o monitoramento e análise critica sejam planejados como parte integrada do processo de gestão de riscos para prover checagem regular e periódica de todo o processo.

As atividades inerentes a este processo devem abranger rodos os aspectos da gestão de riscos com a finalidade de obter informações que possibilitem a melhoria contínua do processo de gerenciamento de riscos; analisar incidentes, mudanças, tendências ou fracassos que demandem alterações no processo de gestão de riscos implantado; monitorar e detectar mudanças nos contextos internos e externos que demandem revisão dos tratamentos de riscos e suas prioridades; identificar riscos emergentes ou que não se encaixem nos critérios inicialmente definidos.

Por fim, a ISO 31.000 sugere que os produtos deste processo sejam documentados, registrados e reportados para todas as partes envolvidas, considerando o plano de comunicação e consulta estabelecido.

Capítulo 4

Estudo de Caso

A Companhia de Saneamento Ambiental do Distrito Federal – CAESB é uma empresa pública de direito privado que foi fundada em 1969 e desenvolve atividades nos diferentes campos do saneamento, em especial no projeto, execução, ampliação, administração, operação e manutenção dos sistemas de abastecimento de água e coleta, tratamento e disposição final de esgoto sanitário em todo o Distrito Federal [43].

Atualmente a CAESB atende cerca de 2,6 milhões de pessoas com serviços de abastecimento de água e 2,5 milhões com serviços de esgotamento sanitário, o que corresponde a 99% e 89% da população instalada no Distrito Federal. No que tange ao esgotamento sanitário, a CAESB trata 100% do esgoto coletado [43].

Os serviços de produção e distribuição de água, bem como coleta e tratamento de esgoto são serviços públicos essenciais [44], de modo que, em se tratando de serviço público essencial, é dever do Poder Público adotar medidas para promoção do saneamento básico [45].

Portanto, não obstante ser uma empresa pública, a CAESB não é titular da prestação desses serviços públicos essenciais, mas tão somente uma concessionária, lhe sendo outorgada a responsabilidade para operação e manutenção das unidades integrantes dos sistemas públicos de abastecimento de água e esgoto do Distrito Federal onde os custos operacionais são arcados mediante cobrança de tarifa [46].

A Agência Reguladora de Águas, Energia e Saneamento do Distrito Federal - ADASA é um órgão de governo e outorgante e reguladora das atividades da CAESB. Em 2011, a ADASA emitiu a Resolução 14 na qual, dentre outras determinações, elencou as atribuições da CAESB para prestação dos serviços de saneamento no âmbito do DF, das quais se destacam [46][47]:

1. O planejamento e a execução das obras e instalações necessárias à regularidade, continuidade, eficiencia, segurança, atualidade, generalidade e universalização dos serviços e modicidade das tarifas;

2. A operação e a manutenção das instalações de captação, adução, tratamento, reservação e distribuição de água;

3. A operação e a manutenção das instalações de coleta, transporte e tratamento do esgoto, e a disposição final dos efluentes líquidos, sólidos e gasosos;

4. A medição dos consumos, o faturamento, a cobrança e a arrecadação de valores;

5. A fiscalização das instalações das unidades usuárias e formas de utilização dos serviços pelos usuários, orientando-os para mudanças e impondo as devidas sanções contratuais.

No âmbito deste trabalho, os itens 4 e 5 acima são os mais relevantes, uma vez que a presente pesquisa visa identificar os riscos ao faturamento da companhia por meio das análises dos dados oriundos da medição de consumo e de fiscalização das unidades usuárias.

Deste modo, faz-se necessário o estabelecimento de diretrizes e metodologia para uma eficiente identificação de riscos ao faturamento, visando garantir que a CAESB possa executar suas atividades primárias com excelência e atingir seus objetivos, em especial a modicidade das tarifas citado no item 1 acima, com o mínimo de adversidade possível.

Neste sentido, este capítulo descreve o trabalho realizado, pautado na revisão da literatura e guiado por *case* semelhante da SANASA, empresa de abastecimento de água e esgotamento sanitário de Campinas/São Paulo, para construção de uma metodologia para identificação de riscos ao faturamento da Caesb baseado no processo de mineração de dados CRISP-DM e alinhado ao processo de gestão de riscos preconizado pela ISO 31000.

4.1 *Benchmarking* Sanasa

Dentre os *cases* pesquisados durante a etapa de revisão da literatura, o trabalho divulgado por Passini e Toledo [11] da Sociedade de Abastecimento e Saneamento, Sanasa, da cidade de Campinas em São Paulo, foi selecionada como *benchmarking* para a parte prática desta pesquisa, tanto pela similaridade dos objetivos, quanto pelo fato de ambas as empresas, Caesb e Sanasa, serem companhias de abastecimento com processos comerciais e obtenção de dados semelhantes.

O *benchmarcking* foi realizado por meio de estudo do trabalho de pesquisa publicado por Passini e Toledo bem como contatos telefônicos e por e-mail para esclarecimento de dúvidas pontuais.

A pretenção durante o estudo de caso foi replicar os caminhos trilhados no trabalho realizado pela Sanasa e evoluir nos pontos elencados pelos autores como passíveis de melhorias.

Cita-se como exemplo, os seguintes pontos destacados pelos autores:

1. Falha na escolha das variáveis relevantes;
2. Falta de conhecimento de estatística;
3. Falta de conhecimento de mineração de dados.

Além destes pontos explicitamente citado pelos autores, identificamos outros pontos que deveriam ser observado antes de iniciar a parte prática do estudo de caso:

1. Sanasa criou um modelo único para todos os clientes;
2. Ausência de novos ciclos de mineração.

A experiência da equipe de ficalização da Caesb demonstra que clientes de diferentes localidades do Distrito Federal, com diferente poder aquisitivo, apresentam características algumas vezes opostas ao cometerem fraudes. Cita-se como exemplo o fato de clientes fraudadores residentes em regiões nobres invariavelmente serem adimplentes com suas contas de água, ao passo que esta característica é oposta nas regiões mais pobres, onde percebe-se como característica comum entre os fraudadores destas regiões a reincidência na inadimplência.

No que tange à realização de novos ciclos de mineração, a revisão da literatura demonstrou ser etapa necessária para refinar o processo de mineração de dados, pois a cada novo ciclo, os resultados obtidos ajudam a identificar necessidade de ajustes no modelo, seja com seleção de novos dados ou exclusão de outros já utilizados, que possibilitam aumentar sua precisão. Cita-se como exemplo, o modelo proposto por Puig [23] que prevê retroalimentação a partir dos resultados obtidos.

Portanto, considerando os três pontos de melhoria elencados pelos autores e os dois supramencionados, o estudo de caso da Caesb foi realizado adotando-se as seguintes evoluções:

Tabela 4.1:

Problemas Identificados	Solução Adotada
Falha na escolha das variáveis relevantes	Criação de grupo de trabalho com participação de representantes especialistas da área de negócio
Falta de conhecimento de estatística	Alocação, no Grupo de Trabalho, de profissional com conhecimento estatístico
Falta de conhecimento de mineração de dados	Capacitação da equipe técnica do Grupo de Trabalho
Modelo único para todos os clientes	Segmentação dos clientes para criação de modelo específico para cada grupo
Ausência de novos ciclos de mineração	Realização de vários ciclos de mineração com retroalimentação dos resultados obtidos

Considerando que as atividades executadas durante o estudo de caso demandaria o envolvimento de profissionais de diferentes áreas da companhia (comercial, fiscalização, TI), em especial de profissionais com profundo conhecimento no negócio (fiscalização e faturamento), para assegurar sua participação efetiva, foi formalizado junto à Diretoria da Caesb um grupo de trabalho nomeando as pessoas chaves de cada área que poderiam contribuir na parte prática desta pesquisa.

No âmbito do grupo de trabalho foram definidas duas equipes:

1. Equipe de negócio responsável por:
    - Passagem do conhecimento do negócio para auxiliar na seleção de variáveis relevantes;
    - Validação dos resultados parciais gerados durante as etapas de seleção, tratamento e processamento de dados; e
    - Verificação, em campo, dos clientes apontados pelo modelo como possível impacto ao faturamento.
2. Equipe técnica responsável por:
    - Captação e integração dos dados necessários;
    - Limpeza, tratamento e processamento dos dados; e
    - Consolidação e apresentação dos resultados.

O grupo de trabalho se reuniu periodicamente (a cada 30 dias) para alinhamento, análise dos resultados parciais e discussão das estratégias a serem adotadas no estudo de caso e a equipe técnica se reuniu semanalmente para por em prática as estratégias definidas, coletar, limpar, tratar e processar os dados e consolidar os resultados para apresentação aos demais membros do grupo.

O modelo gerado, os problemas encontrados e os resultados obtidos são apresentados nas subsessões seguintes.

4.2 Modelo proposto para processo de mineração de dados

O modelo proposto e adotado no estudo de caso apresentado na Figura 4.1 é um processo para mineração de dados adaptado do CRISP-DM que se inicia com a definição de um grupo de trabalho, composto por profissionais da área de negócio com conhecimento dos dados, processos e regras de faturamento e profissionais técnicos com conhecimento em TI e mineração de dados.

A formalização do grupo de trabalho junto a diretoria da Caesb foi fundamental para assegurar a disponibilidade e comprometimento dos profissionais indicados que, de outro modo, em função de suas inúmeras atividades cotidianas, não poderiam dispor de seu tempo para participar de reuniões referentes a este estudo de caso.

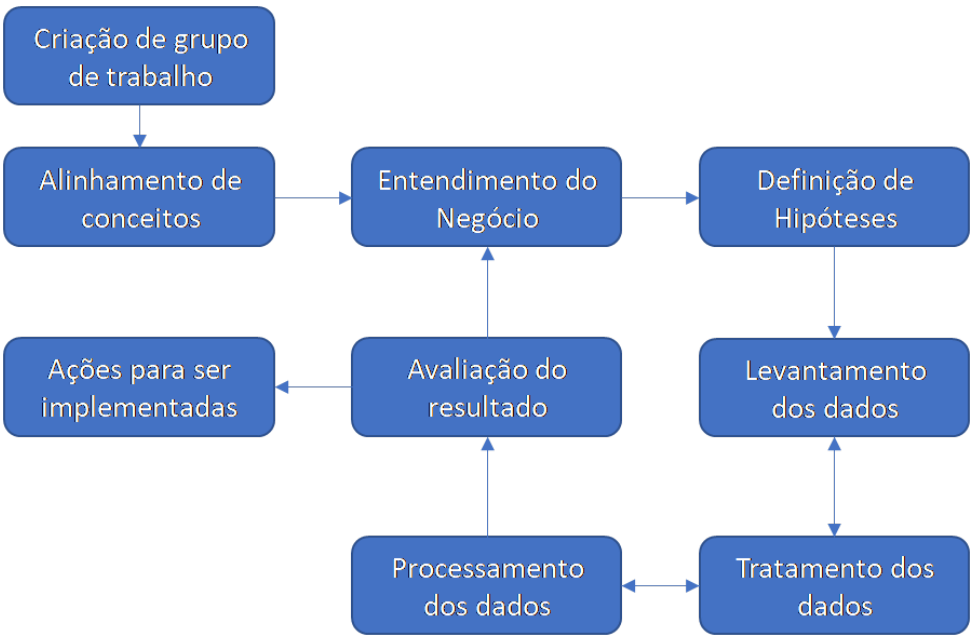


Figura 4.1: Modelo Proposto.

Uma vez definido e formalizado o grupo de trabalho, foi realizado um workshop para alinhamento de conceitos e objetivos do estudo de caso, bem como rápida explanação sobre o que é mineração de dados e resultados alcançado por outras empresas, em especial, pela SANASA, *benchmarking* do presente estudo de caso.

Após esta etapa, iniciou-se a fase cíclica do modelo, com entendimento do negócio, especialmente pela equipe técnica do grupo de trabalho, para construir uma hipótese a ser testada nas etapas posteriores.

Após criação conjunta de uma hipótese, envolvendo todos os membros do gruopo de trabalho, iniciou-se a etapa de levantamento dos dados necessários de serem analisados. Nesta etapa, além de mapear a fonte dos dados (sistemas, tabelas, atributos, etc) e, visando assegurar o entendimento pela equipe técnica, foi explicado pela equipe de negócio os dominios e regras utilizadas na obtenção de cada dado e as variáveis necessárias para os testes das hipóteses criadas foram identificadas.

Para cada hipótese criada, pelo menos um ciclo de levantamento e entendimento, tratamento e processamento dos dados foi executado. Os resultados gerados foram analisados e discutidos com o grupo de trabalho, oportunidade em que a equipe técnica detalhou algumas especificidades do negócio que ficaram evidentes após o processamento dos dados.

Das reuniões de avaliação dos resultados, ações foram encaminhadas para implementação e novos ciclos foram executados a partir de novo detalhamento do negócio e geração de novas hipóteses.

4.2.1 Alinhamento do modelo à ISO 31.000

Conforme mencionado na seção anterior, o modelo aplicado a este estudo de caso é um processo de mineração de dados adaptado do CRISP-DM que, no contexto da cobrança pelos serviços prestados pela Caesb, visa identificar clientes que potencialmente impactam no faturamento.

Neste sentido, importante destacar que este modelo também está alinhado à ISO 31.000, conforme demonstra Figura 4.2, uma vez que se mostra como ferramenta para identificar riscos ao faturamento da Caesb.

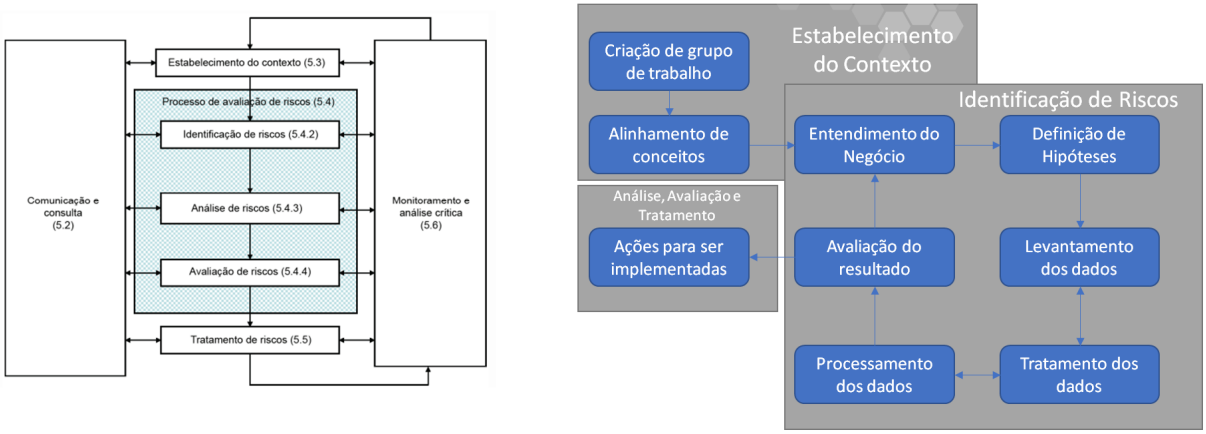


Figura 4.2: Modelo proposto alinhado à ISO 31.000.

Nas etapas de definição do grupo de trabalho, onde profissionais especialistas no negócio e com conhecimento técnico necessário são selecionados e alinhamento de conceitos e entendimento de negócio é realizado, o modelo proposto está desempenhando atividades que permitem a definição do contexto no âmbito da gestão de riscos, ou seja, o contexto em que os riscos serão identificados, analisados, avaliados e tratados.

No presente estudo de caso, este contexto refere-se ao faturamento, onde busca-se identificar fatores de riscos com possível impacto ao faturamento da companhia.

O processo de identificação de riscos do *framework* de gestão de riscos da ISO 31.000 nota-se nas etapas de levantamento, tratamento, processamento e avaliação de resultados do modelo aplicado a este estudo de caso, onde as hipóteses levantadas pelo grupo de trabalho sobre situações de risco ao faturamento são testadas e validadas ou refutadas. Uma vez validada, tem-se identificados os fatores de riscos e os clientes que possivelmente impactam no faturamento da companhia.

Por fim, as ações a serem implementadas após análise dos resultados do teste de hipótese enquadram-se nos processos de análise, avaliação e tratamento de riscos, uma vez que estas ações estão voltadas a mitigar ou eliminar os riscos ao faturamento da Caesb, devendo, antes de serem implementadas, detalhadas e priorizadas, conforme orientação da ISO 31.000.

Importante frisar que a presente pesquisa não tem por pretenção propor um modelo de gestão de riscos ao faturamento, mas tão somente uma ferramenta que permita identificar os clientes que potencialmente impactam no faturamento, portanto, não serão aprofundados temas como a mensuração dos riscos, detalhamento dos critérios e processos para análise, priorização e decisão das formas de tratamento para os riscos identificados. Os esforços foram focados na identificação dos riscos ao faturamento da Caesb usando como ferramenta a mineração de dados baseada no modelo adaptado do CRISP-DM, conforme

exposto na Figura 4.1.

4.3 Hipóteses testadas

Durante os ciclos iniciais de mineração de dados o Grupo de Trabalho, influenciado pelo *case* da Sanasa, focou esforços em uma única abordagem: Impactos no faturamentos causados por fraudes no consumo.

Nesta abordagem, pretendia-se criar um modelo preditivo, utilizando-se o método de classificação com aprendizado supervisionado conforme apresentado em 3.3.2, para identificar possíveis clientes fraudadores.

Para tanto, duas hipóteses foram elaboradas, testadas e refutadas em função da baixa qualidade dos dados, onde os esforços de limpeza e tratamento dos dados não resultaram em uma base consistente o suficiente para possibilitar realizar o treinamento dos algorítimos de aprendizagem supervisionada, etapa necessária para implementação de um modelo preditivo.

Frustados pela impossibilidade de realizar a classificação de clientes fraudadores, condição *sine qua non* para construção de um modelo preditivo e empenhados em alcancar o objetivo do trabalho que é a identificação de possíveis riscos ao faturamento da companhia, uma nova abordagem foi adotada.

Para esta nova abordagem, foi definido o conceito de Par Perfeito, que são os clientes nos quais a Caesb fatura pelos serviços de fornecimento de água e coleta e tratamento de esgoto.

Nesta abordagem, uma hipótese foi elaborada, testada e validada, possibilitando identificar clientes que potencialmente impactam negativamente no faturamento da companhia e clientes que representam oportunidade de aumento na receita.

As seções abaixo apresentam as atividades desempenhadas para testar cada uma das hipóteses elaboradas durante o estudo de caso.

4.3.1 Hipótese 1: Definição de perfil de fraudador por características socioeconômicas

O *case* da Sanasa, utilizado como *benchmarking* para o presente estudo de caso, utilizou-se de apenas 8 variáveis para definição do perfil de clientes fraudadores e criação de um modelo preditivo que, segundo os autores, não atingiu o objetivo.

No *case* da Sanasa, o modelo preditivo considerou:

1. O tipo de fraude de clientes conhecidamente fraudadores;

2. A existência de parcelamento de contas;
3. A situação da ligação de água;
4. A existência de corte no abastecimento;
5. A categoria do imóvel (residência, comércio, indústria, público), onde apenas imóveis residenciais foram utilizados;
6. Média de consumo;
7. Existência de contas retificadas.

Percebe-se que a Sanasa basicamente utilizou variáveis relativas ao relacionamento e consumo dos clientes, o que reforça o relato dos autores que atribuiram a falha na seleção de variáveis como possível responsável pelo não atingimento dos objetivos traçados.

Acreditando que para definição do perfil de fraudadores outras características devem ser consideradas o Grupo de Trabalho do presente estudo de caso levantou a hipótese de que é possível definir o perfil de fraudador por características sociais e econômicas.

Uma vez definida a hipótese, iniciou-se o ciclo de levantamento e processamento de dados previsto no modelo proposto.

Levantamento dos dados

Utilizando técnicas de *brainstorm* as variáveis necessárias para testar a hipótese foram mapeadas e suas fontes de dados identificadas, conforme abaixo:

- **Base cadastral:** Base de dados do sistema comercial da Caesb, denominado GCOM, onde estão armazenados dados de identificação do cliente e cadastro do imóvel.

  Em maio/2019 existiam nesta base 759.832 imóveis e 758.805 clientes cadastrados. A cardinalidade entre a base de imóvel e de clientes é de n:n, ou seja, um imóvel pode ter mais de um cliente associado e um cliente pode possuir mais de um imóvel. No entanto, apenas um cliente consta como responsável financeiro por imóvel, logo, as características sociais do responsável financeiro e que foram consideradas para o teste desta hipótese.

  Da base de cadastro de imóvel, foram selecionadas as seguintes variáveis:

  - Inscricao (integer)
  - IdentificacaoCliente (integer)
  - Logradouro (integer)

  - Data da Ligação de Água (date)
  - Categoria (string)
  - Atividade (string)

  Da base de cadastro de clientes, foram selecionadas as seguintes variáveis:

  - IdentificacaoCliente (integer)
  - Tipo de Cliente no Imovel (string)
  - Tipo de Cliente (string)
  - Clientes Especial(char)
  - Data de nascimento
  - Estado Civil

- **Base geográfica de clientes:** Base de dados que armazena a posição geográfica de cada um dos 759.832 imóveis existentes na base cadastral.

  Da base geográfica de clientes, foram selecionadas as seguintes variáveis:

  - Localizacao do Cliente (geometria)
  - Inscricao (integer)

  Além da base geográfica com o posicionamento dos imóveis, também foi utilizada a base geográfica do Censo realizado pelo IBGE, contendo 4.459 registros para, por meio de análise espacial, identificar a renda *per capta* do setor censitário em que cada imóvel está localizado.

  Da base geográfica do Censo, foram selecionadas as seguintes variáveis:

  - LocalizacaoSetorCensitário (geometria)
  - Regiao Administrativa (string)
  - Renda per capta (float)

- **Base de Irregularidades:** Trata-se de uma base de dados do sistema SICOC, sistema comercial anterior ao atual GCOM, onde as irregularidades identificadas nas ligações de água foram registradas entre 11/01/2001 até 27/09/2013, totalizando 38.074 registros de irregularidades.

  Da base de irregularidades, foram selecionadas as seguintes variáveis:

  - Data da irregularidade (data)
  - Tipo de irregularidade (string)

- **Base de Autuação:** Base de dados do sistema GCOM que armazena as irregularidades identificadas a partir de 30/01/2015, quando foi definido um processo formal de autuação de clientes que cometeram irregularidades nas ligações de água ou esgoto. A partir desta base é possível identificar os clientes que certamente cometeram irregularidades. Esta base contava, até maio/2019, com 11.497 registros.

  Da base de autuação, foram selecionadas as seguintes variáveis:

  - Data da irregularidade (data)
  - Tipo de irregularidade (string)

- **Base de Ordens de serviço:** Base de dados do sistema GCOM que armazena as ordens de serviço para verificação de irregularidades. A partir da análise desta base é possível identificar os clientes que certamente não cometeram irregularidades até a data da vistoria realizada e os clientes que cometeram irregularidades, uma vez que detectada a infração, um registro é gerado na base de autuação a partir da ordem de serviço executada.

  Da base de ordem de serviço, foram selecionadas as seguintes variáveis:

  - Numero da ordem de serviço (string)
  - Inscricao (integer)
  - Data da vistoria (data)

Tratamento dos dados

Na etapa de tratamento de dados, os critérios de seleção foram definidos para cada base de dados levantada, variáveis foram transformadas, registros nulos descartados ou preenchidos com valores que permitissem o processamento dos dados, conforme detalhado abaixo:

Fonte de dado	Critério de Seleção	Variáveis transformadas
Cadastro do Imóvel	Imoveis que constam na base de irregularidade\autuação ou foram realizadas vistorias de irregularidade (ordem de serviço)	- Criada variável que informa se houve mudança de titularidade no imóvel nos últimos 36 meses - Criada variável que informa se houve susbstituição de hidrometro nos ultimos 36 meses - Criada variável que informa se houve desmembramento nos ultimos 36 meses
Cadastro de cliente	Clientes que foram responsáveis financeiros do imóvel no periodo em que foi constatada irregularidade ou realizada vistoria de irregularidade	- Calculado quantidade de dias no imóvel até data da irregularidade - Calculado idade do cliente na data da irregularidade - Classificação do atributo sexo (M = 1 e F = 0) - Classificacao do atributo Cliente especial (1 se não for nulo e 0 se for nulo) - Classificação do atributo estado civil (Solteiro = 1; Casado = 2; Desquitado = 3; Divorciado = 4; Viuvo = 5; Uniao Estavel - 6)
Base de Irregularidades (SICOC)	Irregularidades de água consideradas fraudes*	- Extraido ano da data de irregularidade - Extraído mês da data de irregularidade
Base de Autuacao (GCOM)	Irregularidades de água consideradas fraudes*	- Extraido ano da data de irregularidade - Extraído mês da data de irregularidade
Base de Ordem de Serviço	Ordens de serviço concluídas e do tipo 11312 (Verificacao regularidade servico de agua), 33811 (Denuncia de ligacao clandestina) ou 33822(Vistoria sistemática par reducao de perdas)	

Figura 4.3: Tratamento de dados.

* As bases de dados de Irregularidade (SICOC) e Autuação (GCOM) armazenam diversos tipos de irregularidades de água e esgoto, sendo que muitas delas não caracterizam fraudes por parte do cliente mas que o penalizam com sanções pecuniárias, como por exemplo, descumprimento de orientação por parte da fiscalização.

Por outro lado, a Agência Reguladora (Adasa) definiu um rol taxativo de condutas praticadas pelos clientes que podem ser consideradas como irregularidades pela Caesb para aplicação de sanção. Como os registros na base de dados da Caesb não apresentam as mesmas descrições da relação divulgada pela Adasa, foi necessário realizar o mapeamento e compatibilização das irregularidades registradas nas bases do SICOC e GCOM com as irregulariades definidas pela agência reguladora.

Por este motivo, foi criada uma nova base denominada TipoIrregulariadeAdasa que, combinada com a base de irregularidades (SICOC) e autuação (GCOM) possibilitaram a seleção dos registros que armazenam irregularidades reconhecidas pela Adasa e que caracterizam fraudes cometidas pelo cliente.

Esta atividade demonstra o caráter cíclico do modelo, uma vez que esta necessidade foi identificada na etapa de tratamento de dados, sendo necessário retornar para a etapa anterior de levantamento de dados (mapear as irregularidades definidas pela Adasa) para realizar o tratamento dos dados das bases de irregularidade e autuação.

Processamento dos dados

Após o tratamento dos dados, as bases foram unificadas executando-se atividades de junção, agregação e agrupamento para consolidação de uma base única para mineração dos dados.

Cita-se como exemplo de agregação a inclusão de dados referente a região administrativa e renda per capta do cliente por meio do processamento de analise espacial. A Figura 4.4 ilustra esse processo, onde cada ponto verde representa um cliente e as poligonais dos setores censitários estão delimitadas pelas linhas tracejadas em vermelho.

Figura 4.4: Processamento espacial: Clientes herdando atributos dos setores censitários.

No destaque, os clientes que se localizam dentros da poligonal selecionada receberam os atributos de Renda e região administrativa da poligonal do setor censitário (base geográfica do censo) considerando o posicionamento geográfico do cliente (base geográfica de clientes).

A etapa de processamento dos dados resultou em uma base para mineração com 44.392 registros contendo 40.694 clientes únicos com registros de irregularidades que caracterizam fraudes.

A base de mineração foi utilizada para geração de gráficos que possibilitam melhor entendimento dos dados e perfil de fraudadores considerando as características sócioeconômicas constantes nas variáveis selecionadas.

Avaliação dos resultados

Concluída a etapa de processamento de dados para testar a hipótese 1, os gráficos da Figura 4.5 foram gerados e discutidos pelos membros do grupo de trabalho.

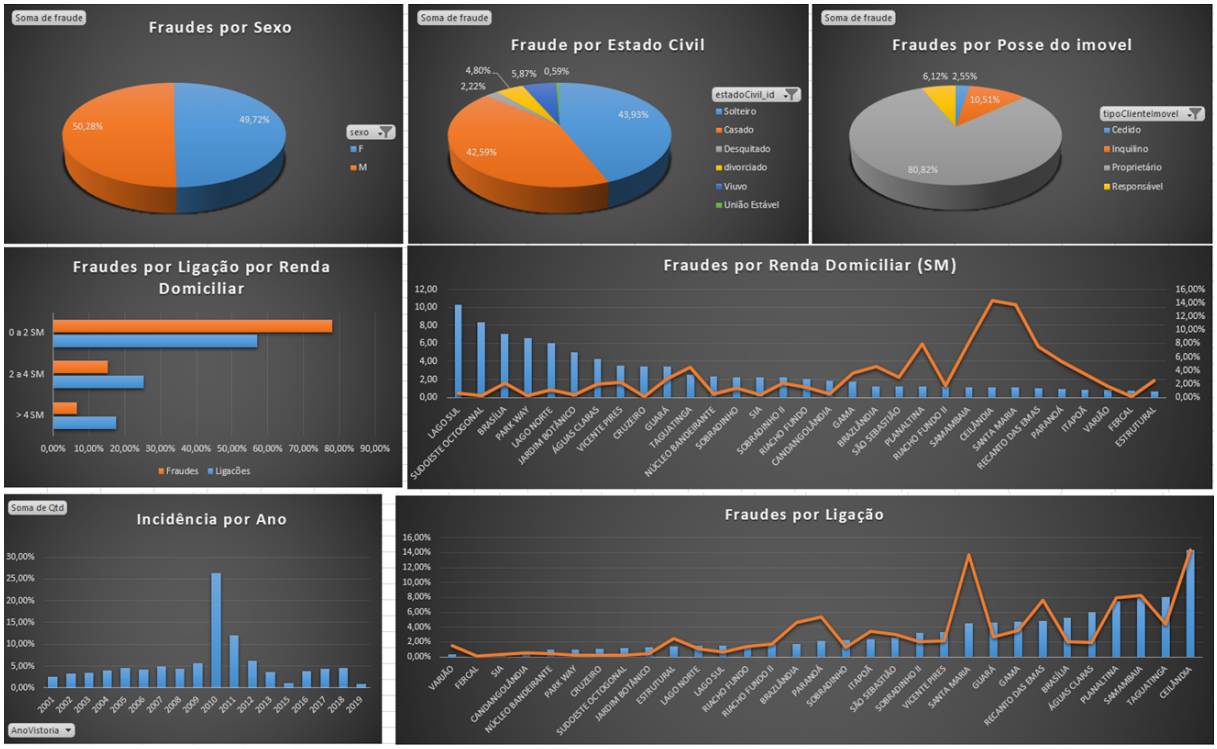


Figura 4.5: Resultado do processamento de dados - Hipótese 1.

O gráfico de Fraudes por Sexo demonstra que, analisando os dados, as irregularidades foram cometidas por homens e mulheres, praticamente na mesma proporção e no que tange ao estado civil, os dados demonstram que as fraudes foram cometidas na mesma proporção por pessoas casadas e solteiras.

Portanto, as características de gênero e estado civil **não são determinantes** para o estabelecer o perfil de cliente fraudador.

O gráfico de fraudes por posse do imóvel demonstra que 80% das fraudes foram cometidas por clientes que são proprietários do imóvel, no entanto, os profissionais da área comercial informaram que **este dado é declarado e não confiável**, pois muitas vezes o proprietário aluga o imóvel e não realiza alteração cadastral junto a Caesb.

Por fim, o gráfico de Fraudes por ligação destaca a região de Santa Maria como localidade com maior registro de fraudes, onde 14% das fraudes registradas foram identificadas nesta região, o que levou a interpretação de que o modelo preditivo deveria se focar nesta região.

Durante a apresentação e discussão dos resultados, profissionais da área de fiscalização explicaram que a grande incidência de fraudes detectadas em 2010 e na região de Santa Maria se justifica pela existência de um contrato de verificação de fraudes, onde praticamente todas as residências de Santa Maria foram vistoriadas e notificadas ao menor sinal

de irregularidade, como por exemplo lacre violado, mesmo aqueles que foram violados devido à oxidação.

Como conclusão da avaliação e discussões dos resultados os profissionais da área de negócio (comercial e fiscalizacão), entenderam que a hipótese 1 foi refutada, uma vez que as variáveis socioeconômicas, além de se mostraram não determinísticas para a definição de um perfil de cliente fraudador, não são confiáveis pela desatualização dos dados e seu caráter auto-declaratório.

Pelo fato da Caesb ter que rever muitas das notificações aplicadas em função do contrato de detecção de fraudes em 2010 e somado a definição de novas regras pela agência reguladora, a metodologia de vistoria de irregularidade foi revista, resultando em 2015 em um novo procedimento e um sistema específico para registro das autuações.

Deste modo, no próximo ciclo do processo de mineração de dados, os profissionais da área de negócio sugeriram utilizar apenas os dados gerados pelo novo procedimento de autuação.

4.3.2 Hipótese 2: Definição de perfil de fraudador pelo padrão de consumo

Após refutada a hipótese 1, ainda focados na abordagem de identificar impactos no faturamento causados por fraudes no consumo, iniciou-se um novo ciclo no processo de mineração de dados elaborando-se uma nova hipótese.

Considerando a experiência da Sanasa que utilizou apenas dados de consumo para construção do modelo preditivo, elaborou-se a hipótese de que é possível definir o perfil de um cliente fraudador por meio de análise do histórico de consumo comparando-se com o consumo da vizinhança.

Levantamento dos dados

Para o teste da segunda hipótese, percebeu-se que as fontes de dados levantadas na primeira hipótese também serão utilizados, excluindo-se as variáveis relativas às características socioeconômicas, sendo necessário acrescentar apenas os dados de consumo.

Os dados de consumo são registrados no sistema comercial (GCOM) na base de leitura, da qual foram selecionadas as seguintes variáveis:

- Inscricao (integer)
- Referencia (integer)
- Volume consumido (integer)

Seguindo a orientação dos profissionais da área de negócio, os dados utilizados para o teste da segunda hipótese deveriam considerar apenas os clientes autuados pelo novo processo de autuação, que foi implantado em 2015.

Com esta restrição, a base de mineração passou a contar com 4.987 registros de fraudes de 4.682 clientes distintos.

Tratamento dos dados

Ao analisar os dados de consumo dos clientes selecionados, percebeu-se que a variável VolumeMedido, que representa o consumo efetivo do cliente, encontravam-se com valores 0 (zero), conforme demonstra a Figura 4.6

inscricao	tipoconsumofaturadodesc	ocorrenciadesc	volumemedido	consumofaturado
281	Mínimo	NÃO INFORMADO	0	10
361	Mínimo	NÃO INFORMADO	0	10
51	Mínimo	Leitura Informada Pelo Usuário	0	10
264	Mínimo	NÃO INFORMADO	0	10
795	Mínimo	Hidrômetro Retirado	0	10
701	Média	Portão Fechado-imóvel Habitado	0	38
825	Mínimo	Portão Fechado-imóvel Habitado	0	10
1295	NÃO INFORMADO	Hidrômetro Retirado	0	0
1287	NÃO INFORMADO	Hidrômetro Retirado	0	0
2151	NÃO INFORMADO	Obstáculo Impedindo a Leitura	0	0
2976	NÃO INFORMADO	Hidrômetro Retirado	0	0
3131	Média	Hidrômetro Enterrado	0	383
1881	NÃO INFORMADO	NÃO INFORMADO	0	0
2259	NÃO INFORMADO	NÃO INFORMADO	0	0
2461	NÃO INFORMADO	NÃO INFORMADO	0	0
2437	NÃO INFORMADO	NÃO INFORMADO	0	0
1911	NÃO INFORMADO	NÃO INFORMADO	0	0
2161	NÃO INFORMADO	Obstáculo Impedindo a Leitura	0	0
1945	NÃO INFORMADO	Obstáculo Impedindo a Leitura	0	0
2879	NÃO INFORMADO	Portão Fechado-imóvel Desabit.	0	0
3018	Média	Portão Fechado-imóvel Habitado	0	27
906	Média	Portão Fechado-imóvel Habitado	0	11
991	Média	Portão Fechado-imóvel Desabit.	0	30
2331	NÃO INFORMADO	Endereço Não Localizado	0	0
2348	NÃO INFORMADO	Endereço Não Localizado	0	0
3077	NÃO INFORMADO	Hidrômetro Retirado	0	0
3123	Média	Obstáculo Impedindo a Leitura	0	85
2666	Média	Hidrômetro Enterrado	0	80
2755	Mínimo	Leitura Informada Pelo Usuário	0	10
2291	NÃO INFORMADO	NÃO INFORMADO	0	0
2305	NÃO INFORMADO	Veículo sem Caixa do Hidrômetro	0	0
2429	NÃO INFORMADO	NÃO INFORMADO	0	0
1856	NÃO INFORMADO	NÃO INFORMADO	0	0

Figura 4.6: Ausência de volume medido.

No processo de faturamento, o volume medido é calculado pela diferença do valor constante no visor do hidrômetro considerando o valor obtido na leitura realizada no mês anterior. No entanto, é frequente a situação em que o leiturista não consegue acessar o hidrômetro, de forma que o VolumeMedido é registrado como 0 e o cliente é cobrado em função do consumo médio.

Na base de leitura, 60,99% dos clientes apresentaram impedimentos de leitura nos últimos 16 meses.

Para testar a hipótese, faz-se necessário conhecer o consumo efetivo do cliente, portanto, para contornar o problema da ausência de dado do volume medido, buscou-se na

base de dados do sistema comercial a tabela que armazena o valor constante no visor do hidrômetro no momento da leitura (LeituraComposição).

A tabela LeituraComposição possui como atributos:

- **Id**: Chave primaria da tabela
- **LeituraId**: Identificação da leitura
- **imovelHidrometroId**: Identificação do hidrômetro
- **DataLeitura**: Data da Leitura
- **LeituraHidrometro**: valor constante no visor do hidrômetro no momento da leitura
- **Medido**: Campo calculado que armazena a diferença entre a Leiturahidrometro atual e a ultima LeituraHidrometro

Destes atributos, apenas alguns foram selecionados e integrados com a base de leitura para facilitar a identificação do imóvel, resultando na tabela da Figura 4.7 que apresenta o valor efetivamente lido no visor do hidrômetro para cada imóvel.

anoMes	imovel_inscricao	leituraHidrometro	dataLeitura	anoMesLeiutura	medido	imovelHidrometro_id	tipoconsumo
201610	117676	11	2016-10-15 00:00:00.000	201610	0	1918853	Medido
201611	117676	21	2016-11-14 00:00:00.000	201611	10	1918853	Medido
201612	117676	29	2016-12-14 00:00:00.000	201612	8	1918853	Medido
201701	117676	31	2017-01-14 00:00:00.000	201701	2	1918853	Medido
201702	117676	37	2017-02-11 00:00:00.000	201702	6	1918853	Medido
201703	117676	48	2017-03-14 00:00:00.000	201703	11	1918853	Medido
201704	117676	55	2017-04-13 00:00:00.000	201704	7	1918853	Medido
201705	117676	63	2017-05-13 00:00:00.000	201705	8	1918853	Medido
201706	117676	71	2017-06-12 00:00:00.000	201706	8	1918853	Medido
201707	117676	77	2017-07-14 00:00:00.000	201707	6	1918853	Medido
201708	117676	83	2017-08-15 00:00:00.000	201708	6	1918853	Medido
201709	117676	90	2017-09-15 00:00:00.000	201709	7	1918853	Medido
201710	117676	98	2017-10-16 00:00:00.000	201710	8	1918853	Medido
201711	117676	104	2017-11-13 00:00:00.000	201711	6	1918853	Medido
201805	117676	142	2018-05-02 00:00:00.000	201805	38	1918853	Medido
201806	117676	147	2018-06-14 00:00:00.000	201806	5	1918853	Medido
201807	117676	156	2018-07-14 00:00:00.000	201807	9	1918853	Medido
201808	117676	168	2018-08-15 00:00:00.000	201808	12	1918853	Medido
201809	117676	179	2018-09-15 00:00:00.000	201809	11	1918853	Medido
201810	117676	197	2018-10-16 00:00:00.000	201810	18	1918853	Medido
201811	117676	209	2018-11-14 00:00:00.000	201811	12	1918853	Medido
201812	117676	209	2018-12-13 00:00:00.000	201812	0	1918853	Medido
201902	117676	221	2019-02-13 00:00:00.000	201902	12	1918853	Medido
201903	117676	228	2019-03-14 00:00:00.000	201903	7	1918853	Medido

Figura 4.7: Amostra de dados de leitura de hidrômetro.

Como se percebe na amostra dos dados da Figura 4.7, a ausência de leitura ocasionada por impedimentos, conforme citado acima, gera *gaps* se considerar as datas de leitura.

Na Figura 4.7, por exemplo, foi realizada uma leitura em Novembro de 2017 (coluna AnoMes = 201711) e outra apenas em Maio de 2018 (coluna AnoMes = 201805) para o imóvel 117676 com o hidrômetro 1918853, gerando um *gap* de leitura de 5 meses.

Um novo *gap* ocorreu entre Dezembro de 2018 e Fevereiro de 2019.

A estratéria adotada para preencher estes *gaps* foi realizar a interpolação dos meses em que efetivamente foi realizada leitura entre os meses de ausência deste dado, seguindo o seguinte procedimento:

1. Identificar, para cada ImovelInscrição, os gaps de leitura, ou seja, os meses em que não foram realizadas leituras no hidrômetro com o mesmo ImovelHidrometroID;
2. Calcular a diferença do campo LeituraHidrometro entre os gaps, os seja, o valor do hidrômetro no primeiro mês após o gap menos o valor do hidrômetro no último mês antes do gap;
3. Dividir a diferença da leituraHidrometro pela quantidade de meses no gap, arredondando para o valor inteiro anterior;
4. Adicionar ao campo medido, o resultado do cálculo do item 3;
5. Para cada registro calculado, alterar o atributo TipoConsumo para "Calculado".

O procedimento acima resulta em uma tabela sem ausência de dados de consumo, conforme Figura 4.8, o que possibilitaria proceder com a análise de padrão de consumo do cliente.

anoMes	imovel_inscricao	leituraHidrometro	dataLeitura	anoMesLeiutura	medido	imovelHidrometro_id	tipoconsumo
201610	117676	11	2016-10-15 00:00:00.000	201610	0	1918853	Medido
201611	117676	21	2016-11-14 00:00:00.000	201611	10	1918853	Medido
201612	117676	29	2016-12-14 00:00:00.000	201612	8	1918853	Medido
201701	117676	31	2017-01-14 00:00:00.000	201701	2	1918853	Medido
201702	117676	37	2017-02-11 00:00:00.000	201702	6	1918853	Medido
201703	117676	48	2017-03-14 00:00:00.000	201703	11	1918853	Medido
201704	117676	55	2017-04-13 00:00:00.000	201704	7	1918853	Medido
201705	117676	63	2017-05-13 00:00:00.000	201705	8	1918853	Medido
201706	117676	71	2017-06-12 00:00:00.000	201706	8	1918853	Medido
201707	117676	77	2017-07-14 00:00:00.000	201707	6	1918853	Medido
201708	117676	83	2017-08-15 00:00:00.000	201708	6	1918853	Medido
201709	117676	90	2017-09-15 00:00:00.000	201709	7	1918853	Medido
201710	117676	98	2017-10-16 00:00:00.000	201710	8	1918853	Medido
201711	117676	104	2017-11-13 00:00:00.000	201711	6	1918853	Medido
201712	117676	110	NULL	NULL	6	1918853	Calculado
201801	117676	116	NULL	NULL	6	1918853	Calculado
201802	117676	122	NULL	NULL	6	1918853	Calculado
201803	117676	128	NULL	NULL	6	1918853	Calculado
201804	117676	134	NULL	NULL	6	1918853	Calculado
201805	117676	142	2018-05-02 00:00:00.000	201805	8	1918853	Medido
201806	117676	147	2018-06-14 00:00:00.000	201806	5	1918853	Medido
201807	117676	156	2018-07-14 00:00:00.000	201807	9	1918853	Medido
201808	117676	168	2018-08-15 00:00:00.000	201808	12	1918853	Medido
201809	117676	179	2018-09-15 00:00:00.000	201809	11	1918853	Medido
201810	117676	197	2018-10-16 00:00:00.000	201810	18	1918853	Medido
201811	117676	209	2018-11-14 00:00:00.000	201811	12	1918853	Medido
201812	117676	209	2018-12-13 00:00:00.000	201812	0	1918853	Medido
201901	NULL	215	NULL	NULL	6	NULL	Calculado
201902	117676	221	2019-02-13 00:00:00.000	201902	6	1918853	Medido
201903	117676	228	2019-03-14 00:00:00.000	201903	7	1918853	Medido

Figura 4.8: Tabela com dados de consumo interpolados.

No entanto, durante o tratamento dos dados para preenchimento das lacunas de medição do consumo, identificou-se as seguintes inconsistências:

- Mais de uma ocorrência, com mesmo valor de leitura, no mesmo dia para o mesmo hidrômetro (Figura 4.9);
- Mais de uma ocorrência, com diferentes valores de leitura, no mesmo dia para o mesmo hidrômetro (Figura 4.10);
- Leituras efetivas com valor inferior à última leitura realizada no mesmo hidrômetro (Figura 4.11);
- Poucas leituras efetivas com valor obtido no visor do hidrômetro igual a zero quando nas leituras anteriores o valor era superior (Figura 4.12);
- Registros de leituras ocorridas no futuro (Figura 4.13).

id	leitura_id	imovelHidrometro_id	dataLeitura	leituraHidrometro	medido	ativo	consumoFaturado	leituraCriada
42489323	42489391	3445	2014-03-07 00:00:00.000	6644	318	1	NULL	0
46112771	46112838	3445	2014-03-07 00:00:00.000	6644	318	1	NULL	NULL
46154003	46154070	3445	2014-04-04 00:00:00.000	6937	293	1	NULL	NULL
43101279	43101346	3445	2014-04-04 00:00:00.000	6937	293	1	NULL	0
43723424	43723492	3445	2014-05-06 00:00:00.000	7263	326	1	NULL	0
46492385	46492452	3445	2014-05-06 00:00:00.000	7263	326	1	NULL	NULL
46492388	46492455	3445	2014-06-04 00:00:00.000	7596	333	1	NULL	NULL
44965628	44965696	3445	2014-06-04 00:00:00.000	7596	333	1	NULL	0
46557472	46557539	3445	2014-07-09 00:00:00.000	7977	381	1	NULL	NULL
45051429	45051497	3445	2014-07-09 00:00:00.000	7977	381	1	NULL	0
46492390	46492457	3445	2014-07-09 00:00:00.000	7977	381	1	NULL	NULL
46492391	46492458	3445	2014-08-05 00:00:00.000	8255	278	1	NULL	NULL
45580141	45580209	3445	2014-08-05 00:00:00.000	8255	278	1	NULL	0
46557469	46557536	3445	2014-09-03 00:00:00.000	8600	345	1	NULL	NULL
46492393	46492460	3445	2014-09-03 00:00:00.000	8600	345	1	NULL	NULL
46363746	46363813	3445	2014-09-03 00:00:00.000	8600	345	1	NULL	0
46819725	46819794	3445	2014-10-03 00:00:00.000	8954	354	1	NULL	0
48783291	48783360	3445	2014-10-03 00:00:00.000	8954	354	1	NULL	NULL
47504052	47504121	3445	2014-11-05 00:00:00.000	9309	355	1	NULL	0

Figura 4.9: Leituras Replicadas.

	id	leitura_id	imovelHidrometro_id	dataLeitura	leituraHidrometro	medido	ativo	consumoFaturado	leituraCriada
13	57157078	57157146	1749992	2016-02-15 00:00:00.000	82	8	1	NULL	0
14	57818623	57818691	1749992	2016-03-15 00:00:00.000	89	7	1	NULL	0
15	58459380	58459448	1749992	2016-04-13 00:00:00.000	95	6	1	NULL	0
16	59109221	59109289	1749992	2016-05-13 00:00:00.000	101	6	1	NULL	0
17	59752989	59753057	1749992	2016-06-14 00:00:00.000	111	0	1	NULL	1
18	60411285	60411353	1749992	2016-07-14 00:00:00.000	121	0	1	NULL	1
19	61040443	61040511	1749992	2016-08-15 00:00:00.000	119	0	1	NULL	0
20	68968928	68968999	1749992	2016-08-15 00:00:00.000	156	2	1	NULL	0
21	61687945	61688014	1749992	2016-09-15 00:00:00.000	125	6	1	NULL	0
22	62368379	62368450	1749992	2016-10-15 00:00:00.000	129	4	1	NULL	0
23	63056922	63056993	1749992	2016-11-16 00:00:00.000	133	4	1	NULL	0
24	63704042	63704113	1749992	2016-12-13 00:00:00.000	136	3	1	NULL	0
25	64360868	64360939	1749992	2017-01-16 00:00:00.000	139	3	1	NULL	0
26	65038312	65038383	1749992	2017-02-13 00:00:00.000	141	2	1	NULL	0
27	65686817	65686888	1749992	2017-03-14 00:00:00.000	143	2	1	NULL	0
28	66340962	66341033	1749992	2017-04-13 00:00:00.000	145	2	1	NULL	0
29	66994754	66994825	1749992	2017-05-15 00:00:00.000	149	4	1	NULL	0
30	67645513	67645584	1749992	2017-06-13 00:00:00.000	151	2	1	NULL	0

Figura 4.10: Leituras Replicadas com valores diferentes.

	imovel_inscricao	leituraHidrometro	dataLeitura	leitura_id	anoMesLeitura	medido	imovelHidrometro_id	tipoconsumo
211	1317	3208	2015-10-14 00:00:00.000	54564216	201510	16	128	Medido
212	1317	3226	2015-11-13 00:00:00.000	55202676	201511	18	128	Medido
213	1317	3242	2015-12-11 00:00:00.000	55841265	201512	16	128	Medido
214	1317	3258	2016-01-13 00:00:00.000	56495873	201601	16	128	Medido
215	1317	9997	2016-02-13 00:00:00.000	57143952	201602	0	1840205	Medido
216	1317	9990	2016-03-14 00:00:00.000	57788732	201603	0	1840205	Medido
217	1317	9983	2016-04-12 00:00:00.000	58430692	201604	0	1840205	Medido
218	1317	9994	2016-05-12 00:00:00.000	59069350	201605	11	1840205	Medido
219	1317	14	2016-06-13 00:00:00.000	59728014	201606	20	1840205	Medido
220	1317	34	2016-07-13 00:00:00.000	60378839	201607	20	1840205	Medido
221	1317	56	2016-08-12 00:00:00.000	61032255	201608	22	1840205	Medido
222	1317	76	2016-09-14 00:00:00.000	61668391	201609	20	1840205	Medido
223	1317	81	2016-10-14 00:00:00.000	62328255	201610	5	1840205	Medido
224	1317	111	2016-11-11 00:00:00.000	63039697	201611	30	1840205	Medido
225	1317	145	2016-12-13 00:00:00.000	63660237	201612	34	1840205	Medido
226	1317	154	2017-01-13 00:00:00.000	64324221	201701	9	1840205	Medido
227	1317	170	2017-02-10 00:00:00.000	65014452	201702	16	1840205	Medido
228	1317	186	2017-03-13 00:00:00.000	65638477	201703	16	1840205	Medido
229	1317	191	2017-04-12 00:00:00.000	66316977	201704	5	1840205	Medido

Figura 4.11: Leituras decrescentes.

imovel_inscricao	leituraHidrometro	dataLeitura	leitura_id	anoMesLeitura	medido	imovelHidrometro_id
795	30	2017-12-11 00:00:00.000	71551178	201712	0	77
795	40	2018-02-08 00:00:00.000	72921357	201802	0	77
795	0	2018-09-14 00:00:00.000	77618297	201809	0	77
795	0	2018-11-13 00:00:00.000	78996003	201811	0	77
795	0	2019-01-10 00:00:00.000	80407255	201901	0	77
795	0	2019-02-11 00:00:00.000	81104333	201902	0	77

Figura 4.12: Poucos registros de leitura.

id	leitura_id	imovelHidrometro_id	dataLeitura	leituraHidrometro	medido
73222973	73223044	1756433	2020-02-24 00:00:00.000	556	11
73222974	73223045	1756524	2020-02-24 00:00:00.000	341	10
73222975	73223046	223360	2020-02-24 00:00:00.000	354	1
73222976	73223047	1865127	2020-02-24 00:00:00.000	178	0
73222977	73223048	1756530	2020-02-24 00:00:00.000	501	0
73222978	73223049	1776969	2020-02-24 00:00:00.000	636	22
73222979	73223050	365063	2020-02-24 00:00:00.000	381	4
73222980	73223051	223358	2020-02-24 00:00:00.000	1248	0
73222981	73223052	223313	2020-02-24 00:00:00.000	1617	0
73222982	73223053	1945104	2020-02-24 00:00:00.000	149	9
73222983	73223054	1834943	2020-02-24 00:00:00.000	140	4
73222984	73223055	1777028	2020-02-24 00:00:00.000	481	12
73222985	73223056	223354	2020-02-24 00:00:00.000	701	6
73222986	73223057	356296	2020-02-24 00:00:00.000	2078	3
73222987	73223058	223317	2020-02-24 00:00:00.000	2017	3
73222988	73223059	1777056	2020-02-24 00:00:00.000	216	4
73222989	73223060	223319	2020-02-24 00:00:00.000	1699	6
73222990	73223061	1777073	2020-02-24 00:00:00.000	357	12
73222991	73223062	1777116	2020-02-24 00:00:00.000	509	15

Figura 4.13: Leituras com data no futuro.

As inconsistências constatadas acima ocorreram por ausência de travas ou regras de validação na entrada dos dados de leitura dos hidrômetros e impossibilitam a definição de um padrão de consumo efetivo, e, consequentemente, testar a hipótese de que é possível identificar perfil de cliente fraudador considerando os dados de consumo.

Portanto, apesar de todos os esforços despendidos durante a etapa de tratamento de dados, utilizando técnicas de interpolação para eliminar os *gaps* causados pela ausência

de leitura, como demonstrado, não foi possível definir o consumo efetivo dos clientes para iniciar as atividades de treinamento dos algorítmos de classificação de um modelo preditivo.

Ainda que se obtivesse sucesso no tratamento e correção destes *gaps* para gerar uma base de dados para treinamento dos algorítimos de classificação do modelo preditivo, a ausência de travas na coleta dos dados torna inviável identificar e tratar mensalmente estes *gaps* na base real do sistema comercial, quando o modelo fosse executado com novos dados de consumo.

Por este motivo, o teste da hipótese 2 foi abortado e passou-se para a segunda abordagem: Impacto no faturamento causado pela ausência de par perfeito.

4.3.3 Hipótese 3: Impacto no faturamento por ausência de Par Perfeito

Diferentes legislações federais bem como a agência reguladora de águas, energia e saneamento básico do Distrito Federal determinam que as edificações urbanas devem estar conectadas às redes de abastecimento de água ou esgotamento sanitário existente, conforme transcrito abaixo:

- Art. 12. É **obrigatória a ligação** de toda construção, considerada habitável, à rede pública de abastecimento de água e aos coletores públicos de esgoto, sempre que existentes. (Lei 5.027/1966 que institui o Código Sanitário do Distrito Federal) grifo próprio.
- Art. 45. Ressalvadas as disposições em contrário das normas do titular, da entidade de regulação e de meio ambiente, toda edificação permanente urbana **será conectada** às redes públicas de abastecimento de água e de esgotamento sanitário disponíveis e sujeita ao pagamento das tarifas e de outros preços públicos decorrentes da conexão e do uso desses serviços. (Lei 11.445/2007 que estabelece as Diretrizes Nacionais para o Saneamento Básico) grifo próprio.
- Art. 6. Excetuados os casos previstos nas normas do titular, da entidade de regulação e de meio ambiente, toda edificação permanente urbana **será conectada** à rede pública de abastecimento de água disponível. (Decreto 7.217/2010 que regulamenta a Lei 11.445/2007) grifo próprio.
- Art. 11. Excetuados os casos previstos nas normas do titular, da entidade de regulação e de meio ambiente, toda edificação permanente urbana **será conectada** à rede pública de esgotamento sanitário disponível.(Decreto 7.217/2010 que regulamenta a Lei 11.445/2007) grifo próprio.

- Art. 31. Toda edificação permanente urbana que esteja em uso e situada em logradouro público que disponha de redes públicas de abastecimento de água e de esgotamento sanitário **deve ser ligada** às mesmas, de acordo com o disposto no Código Sanitário do Distrito Federal – Lei nº 5.027, de 14 de junho de 1966, e na Lei nº 11.445, de 05 de janeiro de 2007. (Resolução 14/2011 – ADASA) grifo próprio.
- Art. 34. Quando o ponto de entrega de água ou de coleta de esgoto estiver a uma distância máxima de 15 (quinze) metros das respectivas redes públicas e não houver necessidade de reforço de capacidade, **o prestador de serviços fica obrigado a executar a ligação** de abastecimento de água ou de esgotamento sanitário nos prazos especificados no Anexo IV, e autorizado a lançar em fatura subsequente o preço do serviço de execução da ligação. (Resolução 14/2011 – ADASA)grifo próprio.

  Nota-se que os dispositivos legais utilizam verbos imperativos que ressaltam a obrigatoriedade das construções permanentes estarem conectadas as redes existentes, de forma que, pela legislação vigente, não é uma faculdade do usuário utilizar-se ou não dos serviços de abastecimento de água e coleta de esgoto fornecidos pela Caesb e é uma obrigação desta proceder com a devida ligação quando o ponto de entrega estiver a uma distância máxima de 15 metros da rede, salvo necessidade de reforço na rede.

Neste contexto, o termo PAR PERFEITO aplica-se às unidades usuárias que estão ligadas às redes de distribuição de água e esgotamento sanitário e gerando receita para a companhia de ambos os serviços prestados, surgindo assim, a hipótese 3 de que existem clientes que violam este conceito e, além de não atentar para legislação em vigor, causam impacto ao faturamento da Caesb.

Levantamento dos dados

Para testar esta hipótese, identificou-se como necessários os seguintes dados do sistema comercial (GCOM)

- Inscricao (integer)
- Situacao da ligaçao de água (string)
- Situacao da ligacao de esgoto (string)

A análise dos dados comercial indicou que mais de 100 mil clientes possuem apenas ligaçao de água, conforme demonstra a Figura 4.14

Figura 4.14: Situação das ligações de água e esgoto.

De fato existem localidades no Distrito Federal que ainda não dispõem de redes coletoras de esgoto, no entanto, para confirmar a hipótese de que algum dos mais de 100 mil clientes violam o conceito de par perfeito, ou seja, estão localizados próximo a rede existente, faz-se necessário realizar análise com dados espaciais, portanto, os seguintes atributos foram selecionados na base de dados geográfica:

- Localizacao do cliente (geometria)
- Inscricao (integer)
- Localizaçao das redes de água (geometria)
- Localizacao das redes de esgoto (geometria)

Tratamento dos dados

Não foi necessário nenhum tratamento aos dados da base comercial ou geográfica para o teste desta hipótese.

Processamento dos dados

Na etapa de processamento, os dados comerciais foram unificados à base geográfica de cliente utilizando como chave de ligação a inscrição do cliente. Deste modo foi possível visualizar em um mapa a localização de cada cliente constante na base comercial. Após a unificação, selecionou-se os clientes que, segundo a base de dados comercial, possuiam apenas ligação de água, conforme demonstra a Figura 4.15

Figura 4.15: Mapa de clientes que possuem apenas ligaçao de água.

Na sequência, acrescentou-se ao mapa os dados geográficos das redes de esgoto para verificar a existência destas redes próximo aos clientes apontados pelo sistema comercial como clientes que possuem apenas ligação de água, conforme Figura 4.16

Figura 4.16: Mapa redes de esgoto.

Avaliação dos resultados

Os pontos azuis no mapa apresentado na Figura 4.16 representam os clientes que, segundo o sistema comercial, estão ligados às redes de água e esgoto, enquanto os pontos vermelho representam os clientes que não estão conectados à rede de esgoto.

As linhas em amarelo, por sua vez, representam as redes de esgotamento sanitário.

Pela combinação dos dados do sistema comercial e dados geográfico de clientes e redes de esgoto, percebe-se a existência de clientes que não estão sendo cobrados pelo serviço de coleta e tratamento de esgoto, mesmo existindo rede coletora próxima ao imóvel, confirmando, assim, a hipótese de que existem clientes que violam o conceito de par perfeito e, portanto, impactam no faturamento da companhia.

4.4 Ferramenta de identificação de riscos ao faturamento

Uma vez definido o conceito de par perfeito e confirmada a hipótese de que a ausência de par perfeito impacta no faturamento da companhia, faz-se necessário identificar quem são, quantos são, onde estão e qual o impacto financeiro causado pelos usuários que não atendem a este conceito.

Para tanto, foi desenvolvida uma ferramenta de identificação de riscos ao faturamento da companhia que envolve a descoberta de usuários que violam o conceito de par perfeito por meio de um modelo de mineração de dados que combina a análise geoespacial de dados do sistema comercial e do cadastro técnico das redes de abastecimento de água e esgotamento sanitário.

O resultado desta mineração de dados é apresentado em um Dashboard Web que consolida a quantidade, localidade, divergências e estimativa do impacto financeiro causado pela violação do par perfeito.

Neste sentido, esta seção apresenta o conceito de geoprocessamento e análise geoespacial, descreve o modelo de mineração de dados implementado após a execução do processo de mineração de dados proposto e apresenta a ferramenta de identificação de riscos ao faturamento que possibilita visualizar, em um mapa no *dashboard web*, o local, a quantidade e o impacto em termos financeiros, causados pelos clientes apontado pelo modelo de mineração de dados como potencial causador de riscos ao faturamento da companhia.

4.4.1 Georreferenciamento e análises geoespaciais

Como se percebe na análise dos resultados da hipótese 3, ela só pode ser confirmada pela análise geoespacial, ainda que visual, da proximidade entre os clientes cujo cadastro

comercial informava haver apenas ligação de água ativa e as redes de esgoto.

Segundo Marques *et al* [48], a análise geoespacial permite a compreensão de fenômenos antrópicos e ambientais, a partir da identificação de padrões e relacionamentos espaciais e mensuração de indicadores.

No entanto, a realização de análises geoespaciais demandam o georreferenciamento prévio dos objetos ou elementos que se deseja analisar.

O georreferenciamento, por sua vez, consiste em posicionar em um mapa, por meio de coordenadas geográficas, edificações, construções, cidades, objetos ou qualquer elemento que se deseje identificar seu posicionamento.

Os elementos georreferenciados são simbolizados por geometrias (linhas, pontos ou polígonos). Por exemplo, a localização de um cliente é representado por um ponto, as redes de água e esgoto são representadas por linhas e os limites das regiões admistrativas do DF são representadas por poligonos.

Para armazenar, consultar, alterar ou manipular dados georreferenciados é necessário utilizar Sistemas de Informações Georreferenciadas - SIG, ou GIS como é mais conhecido pela sua sigla em inglês.

Em um SIG, os dados georreferenciados são representados em camadas onde, além do posicionamento, outros dados relevantes referente ao elemento georreferenciado fica disponível para manipulação.

As camadas podem ser combinadas, conforme ilustra a Figura 4.17 para gerar mapas temáticos, realizar análises geoespaciais ou gerar informações que apoiem na tomada de decisão.

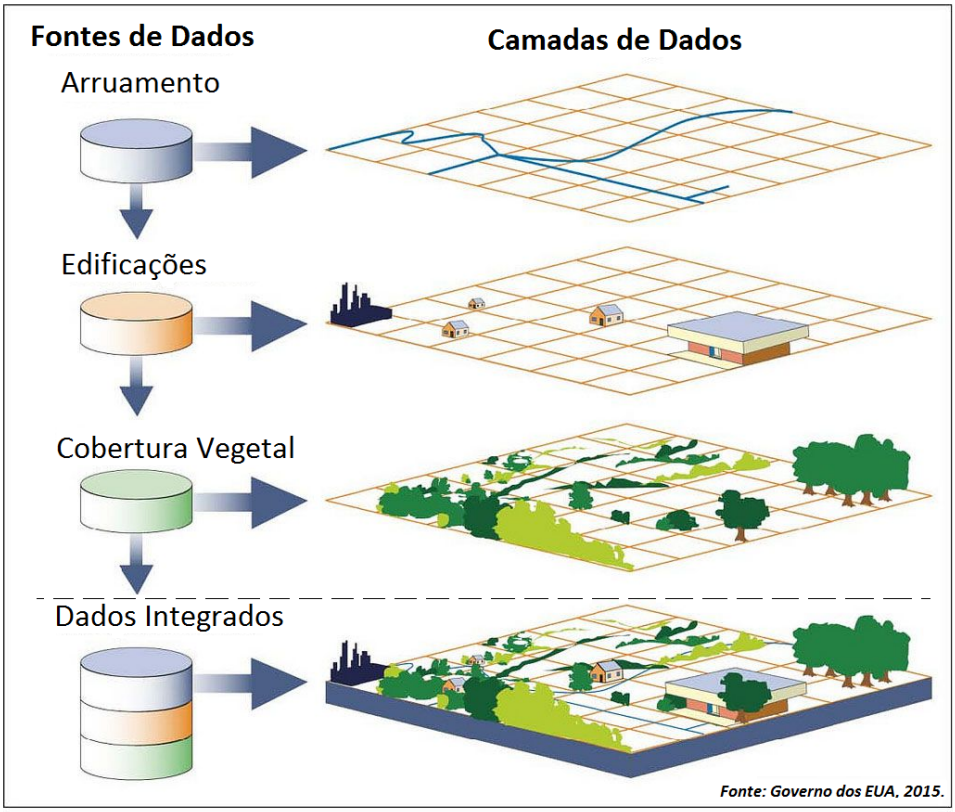


Figura 4.17: Camadas de dados em um SIG.

A Caesb implantou seu SIG corporativo em 2010, onde o cadastro técnico das redes de abastecimento de água e esgotamento sanitário do Distrito Federal, que exisitam apenas em arquivos CAD (*Computer Aided Design*) que era simplesmente uma representação gráfica (desenho) do traçado das redes em relação ao desenho do urbanismo, foi georreferenciado e armazenado em um banco de dados geográfico que, além de dados tabulares, como tipo de rede, material e diametro da tubulação, armazena dados geoespaciais (coordenadas geográficas) permitindo representar o traçado das redes em um mapa.

O georreferenciamento dos clientes, por sua vez, foi realizado em 2012. Atualmente, 100% das redes e clientes da Caesb estão gerreferenciados, de modo que ao sobrepor estas duas camadas (clientes georreferenciados e redes georreferenciadas) conforme demonstrado na Figura 4.16, pode-se confirmar a proximidade entre eles e, por meio da análise geoespacial automatizada, medir a distância e quantificar os clientes que, apesar de próximos as redes de abastecimento ou esgotamento, não estão conectados a alguma delas e, conseguentemente, não estão pagando pelo serviço disponível.

A arquitetura SIG da Caesb conta com uma *suite* de softwares da ESRI que engloba:

- Banco de Dados Espacial ArcSDE versão 10.6.1 instalado em um SGBD Sql Server 2012;
- Aplicação desktop ArcGis Pro versão 2.3;
- Servidor de geosserviço ArcGis Server versão 10.6.1;
- Portal para publicação de mapas Portal for ArcGis versão 10.6.1;
- Operational Dashboards .

Toda esta arquitetura foi utilizada para construção da ferramenta de identificação de riscos ao faturamento da Caesb.

4.4.2 Modelo de mineração de dados

Conforme mencionado por Alcalá [37] *et al*, a mineração de dados revela informações ocultas sobre um grande volume de dados e permite descobrir padrões, tendências e relacionamentos entre conjuntos de dados que de outra forma permaneceriam desconhecidos.

Esta afirmação se comprova no presente estudo de caso, uma vez que, como exposto na subseção anterior, os clientes e redes de abastecimento e esgotamento da Caesb estão georreferenciados desde 2012, no entanto, somente nesta pesquisa onde as atividades de mineração de dados do modelo proposto em 4.1 foram executadas é que se descobriu, se criou o conhecimento de que a análise do relacionamento entre o posicionamento das redes e as unidades consumidoras apontam potencial impacto no faturamento da companhia.

O processo de mineração de dados proposto possibilitou identificar as fontes de dados que devem ser utilizadas, coletar e tratar estes dados para que um ou vários métodos de mineração descritos em 3.3.2 possa ser aplicado na descoberta de conhecimento.

O teste da hipótese 3 confirmou que existem clientes que impactam no faturamento por violar o conceito de par perfeito, portanto, o método de mineração de dados aplicado deve agrupar clientes com características comuns que os distingue daqueles que não violam o conceito de par perfeito.

Deste modo, o modelo de mineração de dados implementado combina dados de situação da ligação de água e esgoto, bem como o valor faturado de cada cliente com dados do cadastro técnico de redes e adota o método de *clusterização* para, por meio de algorítmo de análise geoespacial de proximidade, agrupar clientes com características semelhantes que potencialmente impactam no faturamento da companhia, conforme demonstrado na Figura 4.18. As características de cada grupo é apresentada na próxima seção.

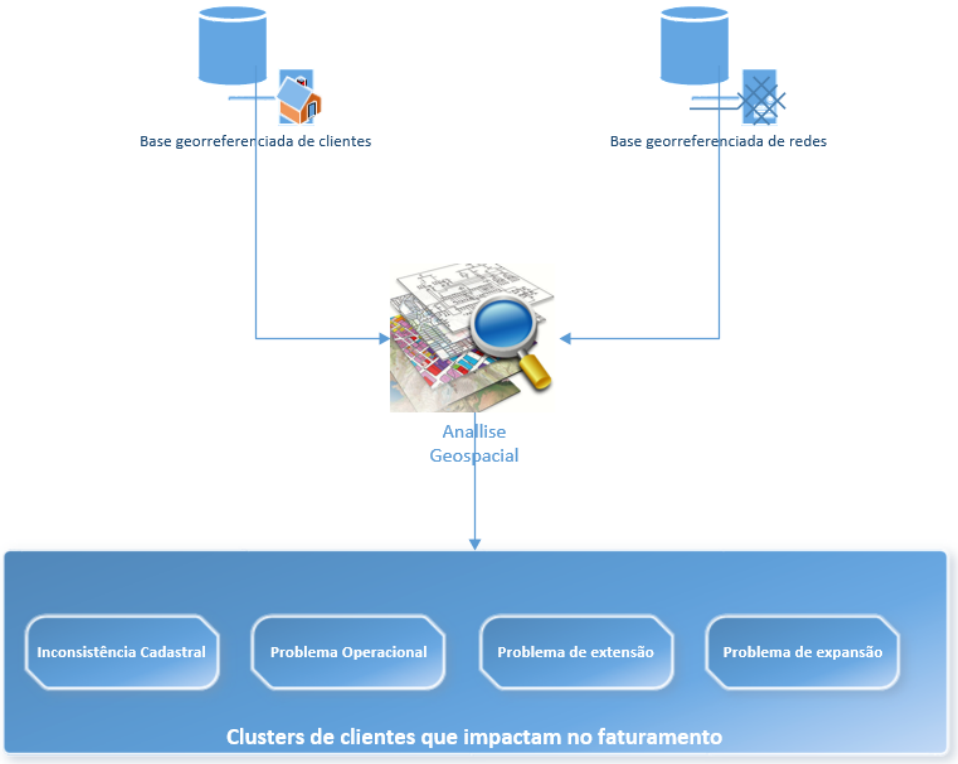


Figura 4.18: Representação do modelo de mineração adotado.

Para estimar o impacto financeiro causado pela violação do conceito de par perfeito, foi considerado o valor cobrado do cliente referente ao consumo de água e multiplicado pelo percentual da tarifa de esgoto informada no cadastro comercial deste cliente.

Na ausência de um percentual da tarifa, foi adotado o valor de 100%, em observância ao disposto no art. 103, inciso I alínea b e inciso II alínea b da resolução 14 da ADASA.

No caso de o cliente ter sido cobrado apenas pela coleta e tratamento de esgoto, a estimativa do impacto financeiro considerou o valor cobrado dividindo-se pelo percentual da tarifa de esgoto informada no cadastro do cliente que violou o conceito de par perfeito, chegando-se assim a estimativa do impacto financeiro causada por aquele cliente ao faturamento dos serviços de água.

O modelo de mineração de dados acima descrito foi implementado em linguagem SQL e agendando no SGBD para ser executado mensalmente de forma automática, após o fechamento comercial.

4.4.3 *Clusters* de pares não perfeitos

O modelo de mineração de dados identificou três situações em que o conceito de par perfeito é violado e uma em que apesar de não ser violado, impacta no faturamento da

companhia. São eles:

1. **Inconsistência Cadastral**, onde a situação de água está inativa enquanto a situação de esgoto está ativa, ou vice-versa, fazendo com que o faturamento ocorra apenas no tipo de ligação ativa.

2. **Problema operacional**, onde a situação da ligação de água ou esgoto é factível ou potencial, ou seja, não está conectado à rede, e existe rede de abastecimento ou esgotamento em um raio de 15 mts da unidade usuária (art. 34 da Resolução 14 da Adasa).

3. **Problema de extensão da rede,** onde a situação da ligação de água ou esgoto é factível ou potencial, havendo outros usuários no mesmo logradouro com situação ativa ou inativa (ou seja, seus vizinhos estão conectados na rede) OU existe rede de abastecimento ou esgotamento entre 15 e 500 mts da unidade usuária com situação factível ou potencial (Art. 35 da Resolução 14 da Adasa).

4. **Problema de expansão da rede**, onde a situação da ligação de água ou esgoto é factível ou potencial e a rede se encontra a mais de 500 mts da unidade usuária do cliente, sendo necessário realizar obra de expansão.

Os problemas 1 e 2 violam o conceito de par perfeito, uma vez que a Caesb não está faturando o serviço prestado para fornecimento de água ou coleta de esgoto apesar do cliente estar conectado em ambas as redes (problema 1) ou pelo fato das redes de abastecimento de água e esgotamento sanitário estarem próximas ao lote do cliente (até 15 mts) e este não estar conectado a alguma delas (problema 2).

O problema 3 viola o conceito de par perfeito pelo fato das redes de abastecimento de água e esgotamento sanitário estarem próximas ao lote do cliente (entre 15 mts e 500 mts) e este não estar conectado a alguma delas.

As legislações vigentes que determinam a obrigatoriedade das construções permanentes estarem conectadas às redes existentes não estipulam uma distância máxima entre o ponto de entrega e as redes de abastecimento de água e esgotamento sanitário, portanto, para este trabalho foi arbitrado a distância de 500 metros para considerar a violação ao conceito de par perfeito.

Esta distância foi arbitrada em função de ser a média dos maiores projetos de extensão de redes realizado pela Gerência de Pequenos Projetos da Caesb, área responsável pelo projeto e orçamentação de extensão de pequenos trechos de rede para fornecimento de água ou coleta de esgoto do cliente, sem que seja necessário realizar reforços nas redes, como instalação de boosters ou elevatórias.

O problema 4, por fim, não viola o conceito de par perfeito pelo fato do cliente estar distante das redes existentes (mais de 500 metros), no entanto, impactam no faturamento da companhia por ser uma oportunidade de aumento de receita, onde serão necessários realizar estudos de viabilidade, projeto e obras para expansão das redes, com investimento de recursos próprios ou de terceiros (Bancos, Governo).

4.4.4 Dashboard de Impacto no Faturamento

A execução mensal do modelo de mineração de dados para identificar possíveis clientes que impactam no faturamento da companhia resulta em uma tabela de banco de dados que consolida a inscrição, o endereço, a categoria e atividade do cliente, o *cluster* que o cliente foi atribuído e o potencial impacto financeiro que este cliente causa ao faturamento da companhia.

Esta tabela é apresentada em um Dashboard Web, por meio do *software Operational Dashboards* da ESRI, fabricante da *suite* de softwares SIG utilizado pela Caesb, conforme Figura 4.19.

Neste *dashboard* os dados quantitativos e qualitativos são apresentados por meio de gráficos de barra e a localização de cada cliente apresentada em um mapa por meio de pontos.

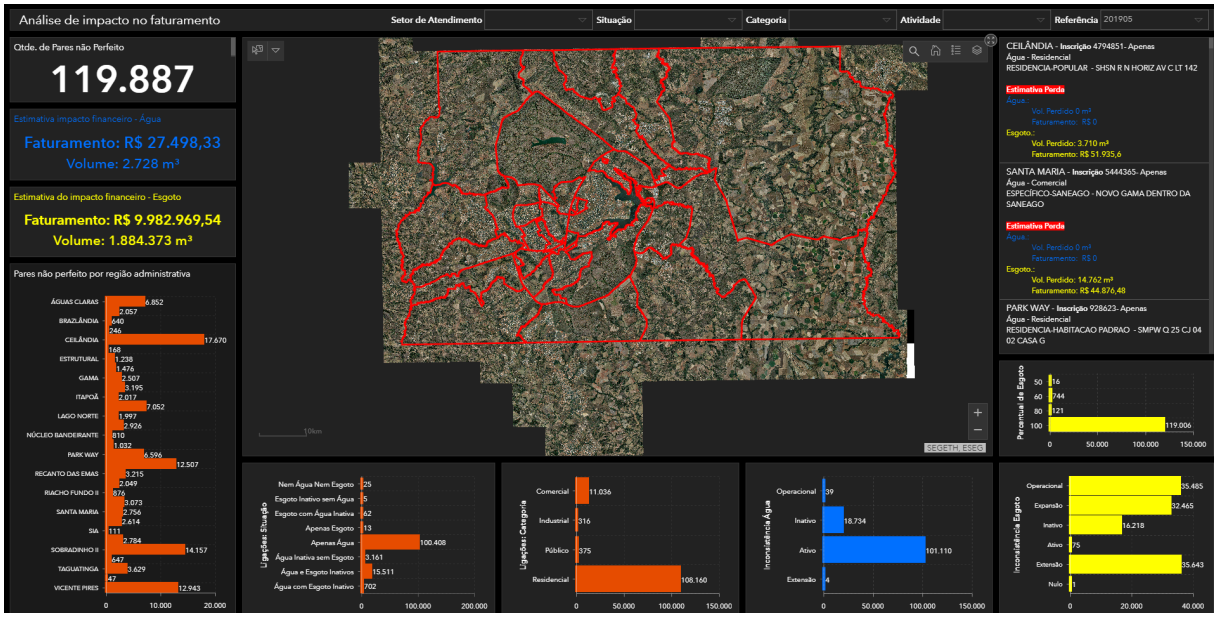


Figura 4.19: Dashboard de impacto no faturamento pela ausência de par perfeito.

O *dashboard* apresenta informações consolidadas, mês a mês, sobre a quantidade, localidade e estimativa do impacto financeiro causado pela ausência dos pares perfeitos de

forma dinâmica, ou seja, os gráficos e informações apresentadas se adequam à visualização no mapa de modo que, ao dar zoom ou navegar no mapa, são apresentadas informações referentes aos usuários existentes na área em visualização.

Clicando nos gráficos de barra, no mapa ou selecionando uma opção nos campos existentes no topo do dashboard é possível realizar filtros pela Região Administrativa, pelo tipo de problema, pela atividade exercida ou categoria do cliente (Residencial, Público, Industrial ou Comercial).

O painel do canto superior direito apresenta detalhes dos usuários que violam o conceito de par perfeito, demonstrando o número da inscrição, o endereço e a estimativa do impacto no faturamento.

Ao clicar em um destes usuários, o mapa aplica um zoom no cliente selecionado, possibilitando identificar sua localização e existência de redes próximas (Figura 4.20).

Figura 4.20: Detalhe de seleção de um cliente.

Os painéis do canto superior esquerdo, por sua vez, apresentam a quantidade de usuários que violam o conceito de par perfeito e a estimativa do impacto financeiro causado por estes usuários no faturamento dos serviços de água e esgoto.

Como mencionado, o dashboard é dinâmico, de forma que estes painéis apresentam o quantitativo e estimativa de valores financeiros referente aos filtros aplicado ou área de visualização no mapa.

Figura 4.21: Painéis de quantidade e impacto financeiro.

Os painéis da Figura 4.21 apresentam os valores referentes à análise realizada após o fechamento comercial do mês de Maio/2019 de todo o Distrito Federal, sem nenhum filtro aplicado de modo que os valores apresentados refere-se a todos os usuários que violam o conceito de par perfeito, considerando os quatro tipos de problemas (inconsistência cadastral, problema operacional, problema de extensão e problema de expansão).

Percebe-se neste painel que em Maio/2019 119.887 clientes causaram impacto no faturamento da companhia na ordem de R$ 10 Milhões de reais, apenas naquele mês.

O impacto no faturamento se divide em duas categorias: Impacto por perda no faturamento e impacto por oportunidade de aumento de receita.

4.4.5 Perda de faturamento

A perda no faturamento se caracteriza pela receita que a Caesb está deixando de obter mensalmente em função de problemas cuja ação saneadora não demandam investimentos e eventuais custos operacionais são arcados pelo cliente, ou seja, não geram despesas para Caesb para serem saneados e possuem retorno imediato.

Se encaixam nesta categoria as divergências cadastrais e problemas operacionais, que em Maio/2019 resultaram em possível perda financeira na ordem de R$ 3.5 Milhões, conforme demonstrado na Figura 4.22.

Figura 4.22: Estimativa de Perda de Faturamento.

O problema de Inconsistência Cadastral foi ocasionado por 764 usuários cuja situação da ligação de água é ativa e esgoto inativa,ou vice-versa. Ou seja,o cliente está conectado na rede de água e esgoto, mas não está sendo cobrado pelos serviços prestados em uma delas, resultando em uma possível perda de faturamento de R$ 84.642,12 em Maio/2019.

O problema operacional, por sua vez, foi ocasionado por 35.499 usuários que estão a menos de 15 metros da rede de água ou esgoto existente, mas não estão sendo cobrados pelo serviço disponível, resultando em uma possível perda de faturamento de R$ 3.428.686,69 em Maio/2019.

4.4.6 Oportunidade de aumento de receita

As oportunidades de aumento de receita se caracterizam pelo aumento de faturamento resultante de obras realizadas pela Companhia para extensão ou expansão das redes para proceder com novas ligações, sendo estas obras precedidas de estudos de viabilidade e projetos que envolve diferentes áreas da companhia. Se encaixam nesta categoria os problemas de extensão e problemas de expansão, que em Maio/2019 resultaram em oportunidade de aumento de receita na ordem de R$ 6.5 Milhões, conforme demonstrado na Figura 4.23.

Figura 4.23: Estimativa de aumento de receita.

O problema de Extensão de Redes foi ocasionado por 35.647 usuários que estão entre 15 e 500 mts da rede de água ou esgoto existente, mas não estão sendo cobrados pelo serviço disponível, resultando em um possível aumento de receita na ordem de R$ 3.418.655,06/Mês (referência Maio/2019) caso passem a ser conectados às redes e faturados pelo serviço prestado.

O problema de Expansão de Redes foi ocasionado por 32.465 usuários que estão a mais de 500 mts da rede de água ou esgoto existente, resultando em um possível aumento de receita na ordem de R$ 3.078.353,47/Mês (referência Maio/2019) caso passem a ser conectados às redes e faturados pelo serviço prestado.

Capítulo 5

Conclusões

O modelo proposto por este trabalho para o processo de mineração de dados aplicado à Caesb possibilitou a criação de uma ferramenta de identificação de riscos aos faturamento que, por meio de um modelo de mineração de dados combinado com sistema de informação geográfica permitiu identificar, quantificar e estimar o impacto financeiro causado pelos clientes que violam o conceito de par perfeito, ou seja, clientes que não geram receita para Caesb dos serviços de fornecimento de água e coleta de esgoto.

Vale ressaltar que a ferramenta desenvolvida não demandou nenhum custo adicional à companhia, uma vez que foi criada utilizando softwares já existentes na empresa e com pessoal próprio, o que potencializa seu custo-benefício.

O conhecimento adquirido durante o Programa de Pós-Graduação em Computação Aplicada - PPCA, foi primordial para o alcance dos resultados, pois como mencionado, os dados e ferramentas utilizados já existiam na companhia há bastante tempo. Portanto o conhecimento dos processos e métodos de mineração de dados obtidos neste mestrado profissional possibilitaram uma iniciativa inédita na Caesb, que é a utilização da mineração de dados por meio de um processo de mineração adaptado para a realidade da companhia e de um modelo que integra diferentes técnicas (clusterização e análise geoespacial) para apontar potenciais clientes que impactam no faturamento da companhia de forma assertiva e precisa.

No que tange ao alcance dos objetivos desta pesquisa, é inquestionável que o objetivo geral foi atingido, uma vez que a ferramenta aponta, individualmente, cada um dos clientes que impactam no faturamento da Caesb, como demonstrado em 4.4, onde a execução do modelo de mineração em Maio/2019 encontrou 119.887 usuários que são pares não perfeitos e uma estimativa de impacto no faturamento de R$ 10.010.467,87 (Dez milhões, dez mil e quatrocentos e sessenta e sete reais e oitenta e sete centavos) naquele mês, aproximadamente R$ 120 Milhões por ano, o que equivale a 7,5% da receita bruta anual da companhia.

Além disto, a ferramenta desenvolvida no âmbito deste trabalho tem aplicação direta nos níveis tático e estratégico da companhia, atendendo a cada um dos objetivos específicos elencados na introdução.

No nível tático, o Dashboard possibilita aos gestores:

1. Monitorar a qualidade do cadastro comercial no que tange a situação das ligações de água e esgoto, que impactam diretamente no faturamento;

2. Acompanhar as ações saneadoras dos problemas operacionais;

3. Realizar análise comparativa, mês a mês, após cada fechamento comercial, da evolução das ações corretivas e eventual crescimento de problemas que impactam no faturamento.

Cita-se como exemplo a comparação dos resultados de Maio/2019 e Junho/2019 apresentada na Figura 5.1.

Figura 5.1: Comparação dos resultados de Maio e Junho de 2019.

Esta tabela demonstra que houve uma redução de 1.660 clientes do grupo de problemas de expansão de redes, no entanto, os grupos de problema operacional e extensão de redes tiveram acrescimo na quantidade de ligações, o que consequentemente, aumentou o impacto financeiro causado no mês de junho/2019.

Ou seja, a solução desenvolvida possibilita monitorar o crescimento dos problemas que causam impacto ao faturamento da companhia e auxiliar ao gestor na priorização de ações saneadoras.

No nível estratégico, por sua vez, o Dashboard apoia na tomada de decisões ao identificar situações e localidades com maior potencial de retorno financeiro pois:

1. Possibilita identificar localidades com maior perda de faturamento para priorizar ações saneadoras;

2. Apoia na análise de viabilidade financeira nos projetos de extensão ou expansão da rede considerando as estimativas de oportunidade de aumento de receita.

Cita-se como exemplo o problema de extensão da rede de esgoto no Lago Norte, ilustrado na Figura 5.2, em que 26 unidades usuárias não estão conectadas às redes existentes a menos de 500 (quinhentos) metros.

Figura 5.2: Problema de extensão de redes no Lago Norte.

A estimativa de aumento de receita para atendimento a essas 26 unidades usuárias é de R$ 9.172,06/mês, cerca de R$ 110 Mil/ano (referência Maio/2019). Segundo a área de pequenos projetos da Caesb, o orçamento estimado para extensão de redes nesta localidade é de R$ 80 mil, logo, caso a Caesb arque com os custos da obra de extensão das redes, o *payback* será de aproximadamente 9 (nove) meses.

Importante lembrar que parte destes custos pode ser cobrado do cliente, uma vez que, conforme o art. 34 da Resolução 14 da Adasa, a Caesb está autorizada a proceder com a cobrança das despesas de ligações cuja rede está localizada até 15 (quinze) metros da unidade usuária.

Além do curto *payback*, o eventual investimento realizado pela Caesb também será revertido na revisão tarifária e contemplado com 8,58% de WACC concedido pela agência reguladora pelo investimento nas obras de extensão da rede.

Por todo o exposto, têm-se de forma inequivoca que a presente pesquisa resultou em uma ferramenta de aplicação prática para identificação de riscos ao faturamento possibi-

litando à Caesb rápido retorno financeiro ao planejar e priorizar de forma proativa ações de vistoria com intuito de minimizar os impactos causados pelos clientes que, apesar de estarem conectados ou próximos às redes existentes, não são faturado pelos serviços prestados.

Soma-se a isto, as oportunidades de expansão de mercado e aumento de receita apontadas pela solução desenvolvida, que além de demonstrar graficamente o potencial de aumento de faturamento por região, auxilia na tomada de decisão de investimentos ao fornecer dados que possibilitam estimar o prazo de *payback* e, assim, calcular o retorno do investimento (ROI).

Referências

[1] Distrito Federal, Governo do: *Plano integrado de enfrentamento à crise hídrica*, 2017. xii, 1, 2

[2] Estatística IBGE, Instituto Brasileiro de Geografia e: *Censo demográfico de 1991. disponível em: https://ww2.ibge.gov.br/home/estatistica/defaultcenso1991.shtm*. 1

[3] Estatística IBGE, Instituto Brasileiro de Geografia e: *Projeção populacional. disponível em: https://www.ibge.gov.br/apps/populacao/projecao/*, 2017. 1

[4] Águas, Energia e Saneamento Básico do Distrito Federal Agência Reguladora de: *Resolução 20 que estabelece o racionamento no abastecimento de água*, 2016. 2

[5] Hamdy, A, Ragab Ragab e Elisa Scarascia-Mugnozza: *Coping with water scarcity: water saving and increasing water productivity*. Irrigation and drainage, 52(1):3–20, 2003. 2

[6] Ezbakhe, Fatine e Agustí Pérez Foguet: *Considering data uncertainty in the water and sanitation sector: application to large number of alternatives and criteria*. Em *EWRA 2017: 10th World Congress on Water Resources and Environment: Athens, Greece: July 5-9, 2017: proceedings book*, páginas 241–248, 2017. 2

[7] Castro Fettermann, Diego de, Kelly Cerqueira Guerra, Aline Patricia Mano e Giuliano de Almeida Marodin: *Uma sistemática para detecção de fraudes em empresas de abastecimento de água*. Interciencia, 40(2), 2015. 2, 4, 5, 23, 24

[8] Brasil, Ministério das Cidades: *Sistema nacional de informações sobre saneamento: Diagnóstico dos serviços de Água e esgoto - 2016*, 2018. `http://www.snis.gov.br/diagnostico-agua-e-esgotos/diagnostico-ae-2016`. 2, 3, 4, 16

[9] Lambert, A. e W. Hirner: *Losses from water supply systems: Standard terminology and recommended performance measures*. The Blue Pages - International Water Association (IWA), 2000. 2, 3, 15

[10] Mutikanga, Harrison E, Saroj K Sharma e Kalanithy Vairavamoorthy: *Assessment of apparent losses in urban water systems*. Water and Environment Journal, 25(3):327–335, 2011. 3, 17

[11] Passini, Silvia Regina Reginato e Carlos Miguel Tobar Toledo: *Mineração de dados para detecção de fraudes em ligações de água*. XI SEMINCO-Seminário de computação, 2002. 5, 23, 25, 30, 37

[12] MARIANO, Ari Melo e Maíra Santos ROCHA: *Revisão da literatura: Apresentação de uma abordagem integradora.* Em *AEDM International Conference–Economy, Business and Uncertainty: Ideas for a European and Mediterranean industrial policy. Reggio Calabria (Italia)*, 2017. 9

[13] Kumar, Sanjiv, Neeta Kumar e Saxena Vivekadhish: *Millennium development goals (mdgs) to sustainable development goals (sdgs): Addressing unfinished agenda and strengthening sustainable development and partnership.* Indian journal of community medicine: official publication of Indian Association of Preventive & Social Medicine, 41(1):1, 2016. 11

[14] Nações Unidas, Organização das: *Objetivos de desenvolvimento do milênio*, Abril 2018. `https://nacoesunidas.org/tema/odm/`. 11

[15] Nações Unidas, Organização das: *Millennium development goals*, Abril 2018. `http://www.un.org/millenniumgoals/`. 11

[16] Nações Unidas, Organização das: *Mdg 7: Ensure environmental sustainability*, Abril 2018. `http://www.mdgmonitor.org/mdg-7-ensure-environmental-sustainability/`. 12, 13

[17] Nações Unidas, Organização das: *Goal 7: Ensure environmental sustainability*, Abril 2018. `http://www.un.org/millenniumgoals/environ.shtml`. 12

[18] Desenvolvimento, Programa das Nações Unidas para o: *Água é vida*, Abril 2018. `http://www.aguaevida.net.br`. 13

[19] Nações Unidas, Organização das: *17 objetivos para o desenvolvimento sustentável*, Abril 2018. `https://nacoesunidas.org/pos2015/`. 13

[20] Naçoes Unidas, Organização das: *Water and sustainable development, from vision to action*, 2017. 13, 14, 15

[21] Lambert, AO: *International report: water losses management and techniques.* Water Science and Technology: Water Supply, 2(4):1–20, 2002. 15, 16, 17

[22] Cole, Megan J, Richard M Bailey, James DS Cullis e Mark G New: *Water for sustainable development in the berg water management area, south africa.* South African Journal of Science, 114(3-4):40–49, 2018. 15

[23] Coma-Puig, Bernat, Josep Carmona, Ricard Gavalda, Santiago Alcoverro e Victor Martin: *Fraud detection in energy consumption: A supervised approach.* Em *Data Science and Advanced Analytics (DSAA), 2016 IEEE International Conference on*, páginas 120–129. IEEE, 2016. 19, 20, 38

[24] Ahmad, Tanveer, Huanxin Chen, Jiangyu Wang e Yabin Guo: *Review of various modeling techniques for the detection of electricity theft in smart grid environment.* Renewable and Sustainable Energy Reviews, 2017. 20

[25] Jan, Chyan long: *An effective financial statements fraud detection model for the sustainable development of financial markets: Evidence from taiwan.* Sustainability, 10(2):513, 2018. 21

[26] Ngai, EWT, Yong Hu, YH Wong, Yijun Chen e Xin Sun: *The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature.* Decision Support Systems, 50(3):559–569, 2011. 21, 29, 30

[27] Humaid, Eyad H e Tawfeeg Barhoum: *Water consumption financial fraud detection: a model based on rule induction.* Em *Information and Communication Technology (PICICT), 2013 Palestinian International Conference on*, páginas 115–120. IEEE, 2013. 21, 23

[28] Randhawa, Kuldeep, Chu Kiong Loo, Manjeevan Seera, Chee Peng Lim e Asoke K Nandi: *Credit card fraud detection using adaboost and majority voting.* IEEE ACCESS, 6:14277–14284, 2018. 22

[29] Carneiro, Nuno, Gonçalo Figueira e Miguel Costa: *A data mining based system for credit-card fraud detection in e-tail.* Decision Support Systems, 95:91–101, 2017. 22, 30

[30] Kho, John Richard D e Larry A Vea: *Credit card fraud detection based on transaction behavior.* Em *Region 10 Conference, TENCON 2017-2017 IEEE*, páginas 1880–884. IEEE, 2017. 22

[31] Verma, Aayushi, Anu Taneja e Anuja Arora: *Fraud detection and frequent pattern matching in insurance claims using data mining techniques.* Em *Contemporary Computing (IC3), 2017 Tenth International Conference on*, páginas 1–7. IEEE, 2017. 22

[32] Hillerman, Tiago, João Carlos F Souza, Ana Carla B Reis e Rommel N Carvalho: *Applying clustering and ahp methods for evaluating suspect healthcare claims.* Journal of Computational Science, 19:97–111, 2017. 22

[33] Monedero, Iñigo, Félix Biscarri, Juan I Guerrero, Moisés Roldán e Carlos León: *An approach to detection of tampering in water meters.* Procedia Computer Science, 60:413–421, 2015. 23, 24

[34] Candelieri, Antonio: *Clustering and support vector regression for water demand forecasting and anomaly detection.* Water, 9(3):224, 2017. 24

[35] Agrawal, Rakesh, Tomasz Imielinski e Arun Swami: *Database mining: A performance perspective.* IEEE transactions on knowledge and data engineering, 5(6):914–925, 1993. 25, 26, 29, 30

[36] Han, Jiawei, Shojiro Nishio, Hiroyuki Kawano e Wei Wang: *Generalization-based data mining in object-oriented databases using an object cube model.* Data & Knowledge Engineering, 25(1-2):55–97, 1998. 25

[37] Alcalá, Rafael, María José Gacto e Jesús Alcalá-Fdez: *Evolutionary data mining and applications: A revision on the most cited papers from the last 10 years (2007–2017).* Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 8(2):e1239, 2018. 25, 63

[38] Azevedo, Ana Isabel Rojão Lourenço e Manuel Filipe Santos: *Kdd, semma and crisp-dm: a parallel overview.* IADS-DM, 2008. 26, 27, 28

[39] Fayyad, Usama, Gregory Piatetsky-Shapiro e Padhraic Smyth: *From data mining to knowledge discovery in databases.* AI magazine, 17(3):37, 1996. 26

[40] Purdy, Grant: *Iso 31000: 2009—setting a new standard for risk management.* Risk Analysis: An International Journal, 30(6):881–886, 2010. 31

[41] Standardization ISO, International Organization for: *Iso 31.000: Gestão de riscos - princípios e diretrizes*, 2009. 31, 32

[42] Governança Corporativa, Instituto Brasileiro de: *Guia de orientação para o gerenciamento de riscos corporativos*, 2007. 33

[43] Distrito Federal, CAESB Companhia de Saneamento Ambiental do: *Relatório anual de administração 2018*, 2018. `https://www.caesb.df.gov.br/images/arquivos_pdf/RelatoriodaAdministracao_2018v1.pdf`. 36

[44] Brasil: *Decreto 7.217, de 21 de junho de 2010. regula a lei 11.445 de 5 de janeiro de 2007*, 2010. 36

[45] Brasil: *Lei 5.027, de 14 de junho de 1966. institui o código sanitário do distrito federal*, 1966. 36

[46] Águas, Energia e Saneamento do DF Agência Reguladora de: *Contrato de concessão 001/2016*, 2006. 36

[47] Águas, Energia e Saneamento do DF Agência Reguladora de: *Resolução 14. estabelece condições da prestação e utilização dos serviços públicos de abastecimento de água e esgotamento sanitário do df*, 2006. 36

[48] Marques, Mara Lúcia, Maurício Corégio Silva e Danilo Mangaba de Camargo: *Análise geoespacial no mapeamento da vulnerabilidade socioambiental em campinas, sp.* Revista Brasileira de Cartografia, 69(9), 2017. 61